DOCUMENT RESUME

ED 309 195 TM 013 681

AUTHOR Tallmadge, G. Kasten; Wood, Christine T.
TITLE ESEA Title I Evaluation and Reporting System: User's Guide.
INSTITUTION RMC Research Corp., Mountain View, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, DC. Office of Planning, Budgeting, and Evaluation.
PUB DATE Oct 76
CONTRACT HEW-300-76-0316
NOTE 79p.
PUB TYPE Guides - General (050)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Gains; *Compensatory Education; Control Groups; Cost Effectiveness; Disadvantaged Youth; Elementary School Students; Elementary Secondary Education; *Evaluation Methods; Inservice Teacher Education; *Norm Referenced Tests; Parent Conferences; *Pretests Posttests; *Program Effectiveness; Program Evaluation; Regression (Statistics); Secondary School Students
IDENTIFIERS Normal Curve Equivalent Scores; *Title I Evaluation and Reporting System

ABSTRACT

This guidebook concentrates primarily on describing the impact-assessment component of the Elementary and Secondary Education Act (ESEA) Title I evaluation and reporting system for users of the system. Three general evaluation models are presented, along with implementation information for each. The first model, a norm-referenced design, may be implemented using norm-referenced or non-normed tests; if norm-referenced tests are chosen, a pretest-posttest design is required. The second model, a control group design, calls for either a normed or non-normed test to be given to treatment and control groups within a pretest-posttest design; however, if a non-normed test is used, a nationally normed test must also be administered to the treatment group sometime during the school year. The third model, a special regression design, can also use norm-referenced or non-normed tests within a pretest-posttest design. The models all focus on obtaining as clear and unambiguous an answer as possible to the question of how much more the students learned by participating in the Title I project than they would have learned without it. Regardless of the model used, cognitive achievement gains are expressed in terms of a common metric--the normal curve equivalent. Ways of measuring gains in normal curve equivalents are discussed. Information is provided on the collection of data pertinent to five other aspects: (1) participation of services by type to public and non-public school students; (2) Parent Advisory Councils; (3) programming costs; (4) program personnel; and (5) training for teachers and teaching aides. Seven figures and one data table are included. (TJH)

USER'S GUIDE

ED309195

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ESEA TITLE 1
EVALUATION & REPORTING SYSTEM

TM 013681

BEST COPY AVAILABLE

USER'S GUIDE

ESEA Title I
Evaluation and Reporting System

G. Kasten Tallmadge
Christine T. Wood

October 1976

Prepared for

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education/Office of Planning, Budgeting, and Evaluation

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

RMC Research Corporation
Mountain View, California

TABLE OF CONTENTS

		Page
TABLE OF CONTENTS		iii
LIST OF FIGURES AND TABLES		iii
ACKNOWLEDGEMENTS		v
I.	OVERVIEW OF THE SYSTEM	1
II.	CHOOSING AN EVALUATION MODEL	19
III.	IMPLEMENTATION OF MODEL A1	37
IV.	IMPLEMENTATION OF MODEL A2	43
V.	IMPLEMENTATION OF MODELS B1 AND B2	47
VI.	IMPLEMENTATION OF MODELS C1 AND C2	54
VII.	ADDITIONAL CONSIDERATIONS	64
REFERENCES		73

LIST OF FIGURES AND TABLES

Figure 1.	Area under the normal curve divided into NCEs, percentiles, stanines, and standard deviations	9
Figure 2.	Decision tree for selecting an evaluation model	20
Figure 3.	Scatter plot of 200 pairs of pretest and posttest scores	55
Figure 4.	A regression line passing through th idpoints of the posttest score distributions	56
Figure 5.	The regression line for a truncated distribution	58
Figure 6.	Treatment and comparison group regression lines	58
Figure 7.	Sample format for a testing schedule	65
Table 1.	The Normal Curve Equivalent Corresponding to Each Tenth of a Percentile	11

ACKNOWLEDGEMENTS

The documentation of the ESEA Title I evaluation and reporting system reflected by this User's Guide represents but a small fraction of the effort that went into the initial development and several subsequent revisions of the system itself. Many people contributed to the total effort. The authors are particularly indebted to other members of the RMC staff including especially Donald P. Horst, Alexander B. Cannara, Alice Chiang, Barbara L. Bessey, Joanne L. Binkley, Barbara M. Fagan, Nona N. Gamel, Richard T. Johnson, Sonia A. Johnson, Lawrence D. Rosen, Lawrence I. Weiner, Liz Cameron, and Marsha McLean.

Much assistance was provided by Janice K. Anderson of U.S.O.E.'s Office of Planning, Budgeting, and Evaluation and by her predecessor as Project Officer, George Mayeske. Both made valuable technical inputs, contributed moral support when we needed it, and significantly smoothed our working relationships with State and local education agencies. They deserve much credit for all that has been accomplished to date.

We are also indebted to our Policy Advisory Panel members; Gary Brummitt, Nancy Burton, Richard Faunce, Constance Gomes, Weldon Idol, Rosella Lipson, Rose Mary Loving, Vincent Madden, Andrew Nutt, Wallace Roby, Paul Rost, Daniel Schooley, and Rickey Silberman. We owe a similar debt of gratitude to the members of our Technical Advisory Group, Donald T. Campbell, Gene V. Glass, Frederic M. Lord, Anthony J. Nitko, Charles S. Reichardt (who represented Donald Campbell at a panel meeting in July 1975), and Richard E. Snow.

Finally we want to thank the many SEA and LEA representatives with whom we worked during the past year. Too numerous to name, they typically put in long hours working with the RMC site visiting teams. Many improvements in the system have resulted from their efforts.

G. Kasten Tallmadge
Christine T. Wood

I. OVERVIEW OF THE SYSTEM

The objective of the Title I evaluation and reporting system is to provide meaningful, comparable information about Title I projects at the school building, school district, State and Federal levels. Data are collected and/or aggregated on six topics: (a) participation, (b) parent advisory councils (PACs), (c) personnel, (d) training, (e) cost, and (f) impact.

The Title I project is the basic unit of analysis, and summary information about individual projects is maintained at all levels of the system. Thus, project characteristics such as hours per week of student exposure to the project, total project hours, instructor-to-student ratio, and cost per student can be related to project gain. This feature enables the identification of promising educational practices for further investigation and possible dissemination. A project is conceived as a set of methods, materials, personnel, and activities that define an instructional treatment which is judged to be uniform for all those it serves. A Title I project may be clearly defined by the funding application, but where a single application results in the funding of several qualitatively different instructional treatments at different sites, the term project is reserved for the individual treatments and not for the funded composite.

Presently this system requires that impact data be collected only from projects serving the basic skill areas (broadly defined to include arithmetic, mathematics, reading, language arts, and communication). Any one of three basic evaluation designs may be selected for the impact assessment. These designs are: Model A, the Norm-Referenced Model; Model B, the Control Group Model; and Model C, the Special Regression Model. Further flexibility is afforded in that each design has variations to accommodate the use of either normed achievement tests (Models A1, B1, and C1) or tests for which normative data are not available (Models A2, B2, and C2). The selection of the

evaluation design and the test instrument can be made at the State, LEA, or project levels.

The three designs included in the evaluation and reporting system are generally acknowledged to be technically sound by the professional community and were selected because they appear feasible to implement in actual school settings. If properly implemented, each will yield valid, comparable, and interpretable results. The focus of all the models is to obtain as clear and unambiguous an answer as possible to the question, "How much more did pupils learn by participating in the Title I project than they would have learned without it?" Thus, for each model, the treatment effect is defined as the difference between observed post-treatment performance (posttest scores) and an estimate of what that performance would have been without the Title I project. The models differ only in the manner in which this no-treatment expectation is generated.

Regardless of the model used, cognitive achievement gains are expressed in terms of a common metric, the Normal Curve Equivalent (NCE). An NCE is a normalized standard score that has been linearly transformed to match the percentile distribution at values of 1, 50, and 99. The scale has a mean of 50 and a standard deviation of 21.06. A more detailed description of NCEs is provided at the end of this chapter.

General Information

This guidebook concentrates primarily on describing the impact-assessment component of the evaluation and reporting system. However, information on five other topics (referred to in the system as General, rather than Impact information) is collected as well. Each of these topics is briefly described below. In all cases, information is first collected at the building level (although the forms need not be filled out by building personnel) and subsequently aggregated at the LEA and State levels.

Participation. The information requested includes (a) an unduplicated count of the number of public and nonpublic school children served by Title I at each grade level, plus a total count by grade and across grades, and (b) a duplicated count of the number of children receiving specified types of services.

Parent Advisory Councils. Information is called for on the number of buildings and districts having PACs and the number of PACs having a majority of members who are parents of Title I students. Listings of the types of activities undertaken by the PACs and the types of assistance given to them by the LEAs are also requested.

Cost. Information is sought concerning the total cost of all projects, the part of the total that comes from Title I funds, and the part that comes from other supplementary funds.

Personnel. The numbers (full-time equivalents) of personnel employed by the Title I program are called for, broken down according to type.

Training. The number of projects that provide various types of training for teachers and/or aides is requested as well as the number of projects which provide different amounts of such training.

Impact Information

The Basic Evaluation Scheme. The three evaluation models use a common definition of the treatment effect in measuring the cognitive impact of Title I projects. The gain attributed to the treatment is defined as the difference between the treatment group's performance on a post-treatment test and an estimate of what performance on the same test would have been had the group not received the treatment. In other words, the project's impact is the observed post-treatment performance minus the expected no-treatment performance. This relationship can be expressed as follows:

$$\text{Treatment Effect} = \begin{bmatrix} \text{Observed} \\ \text{Post-Treatment} \\ \text{Performance} \end{bmatrix} - \begin{bmatrix} \text{Expected} \\ \text{No-Treatment} \\ \text{Performance} \end{bmatrix}. \qquad (1)$$

The observed post-treatment performance is always the mean or median posttest score of the treatment group. The no-treatment expectation is derived through implementation of one of the evaluation models. The norm-referenced design generates a no-treatment expectation from the assumption that, without the special project, the treatment group would maintain its status relative to a national or local norm group from pretest to posttest. The control group model utilizes the posttest (or adjusted posttest) scores of a control group as the no-treatment expectation. In the special regression design, the no-treatment expectation is derived by entering the treatment group's mean pretest score in the comparison group's post-on-pretest regression equation.

The sections that follow provide overviews of each model and outline the basic assumptions underlying them. The methods used to obtain the observed posttest performance and the expected no-treatment performance in terms of percentiles, raw scores, or standard scores are explained. The last section of this chapter deals with the common metric, Normal Curve Equivalents. Characteristics of the scale are described as well as the procedure used to measure gain in NCE units.

Model A: Norm-Referenced Design. Model A can be implemented with either normed tests (Model A1) or non-normed tests (Model A2). Model A1 requires that the treatment group be both pre- and posttested with a nationally or locally normed achievement test.

When tests with national norms are used, the no-treatment expectation is found by determining the percentile status of the treatment group at pretest time. It is assumed that, without the Title I treatment, the status of the group at posttest time would be the same as it was at pretest time. Thus the group's pretest percentile becomes

the expected no-treatment posttest performance. The observed post-treatment performance is simply the percentile rank corresponding to the group's mean posttest score. If the group's posttest status is higher than the no-treatment expectation, the assumption is made that the improvement resulted from participation in the Title I project.

When tests with local but no national norms are used, it is necessary to administer a test with national norms as well. The no-treatment expectation is derived from the local norms, following the same procedure just described. Data from the nationally normed test are needed only to convert the measured gain into NCE units. (NCE conversions are discussed in a later section of this chapter.)

The norm-referenced model may also be implemented using non-normed tests (Model A2). A nationally normed test, however, must be given at pretest time in addition to the non-normed test. The model involves an equipercentile equating of the scores on the normed and non-normed tests so that the norms may be used to derive the observed posttest performance. In essence, the procedure is as follows. The median pretest standard score on the normed test is determined. The pretest percentile corresponding to this test is then read out of the pretest norms table. This percentile constitutes the no-treatment expectation. Using the score equivalency data developed at pretest time, the median posttest score on the non-normed test is next converted to its normed test counterpart. This figure, in turn, is converted to a percentile using the posttest norms table. The percentile derived in this manner is the observed post-treatment performance indicator.

In the norm-referenced model, all pre- and posttesting (normed and non-normed) must be done at times which are close to empirical normative data points for the normed test. For Model A2 this means that the non-normed test must be given near the dates when normative data were collected for the normed test.

Model B: Control Group Design. Model B is a control group design which requires that either a normed or non-normed test be given to treatment and control groups at pretest and posttest times. If a non-normed test is administered pre and post, a nationally normed test must also be given sometime during the school year to the treatment group. These normed test scores are used only to convert the measure of gain into NCEs.

Between pre- and posttesting, the educational experiences of children in the control group should be similar to those of the treatment group children with the single exception that the control group does not participate in the Title I project. The control group's posttest percentile (Model B1) or mean raw score (Model B2) is the no-treatment expectation. The treatment group's percentile or mean raw score on the posttest forms the observed post-treatment performance. When the treatment group's performance is superior to the control group's, it can be assumed that the project was effective.

The pretest scores of the groups are used only to verify their pre-treatment comparability, or to quantify the initial difference between the groups. Two statistical techniques are offered to adjust the posttest score difference for whatever pretest inequality may exist. Where groups were formed by random assignment of children drawn from a single population or where two pre-existing groups are enough alike to be considered random samples from a single population, the covariance method of adjustment is used. Where small systematic differences are known or assumed to exist between treatment and control groups (in other words when the groups are best considered samples from different populations), the principal-axis or standardized-gain-score method of adjustment is used.

The accuracy of the measure of project impact depends on the suitability of the control group. Ideally, the treatment and control groups should be equivalent on all educationally relevant dimensions

such as socioeconomic status, race, sex, etc. Small between-group differences can be adquately handled through the appropriate statistical adjustment procedure. Large differences are likely to invalidate the model. The model permits the use of different tests or different levels of the same test for pretest and posttest, and testing times need not coincide with norming dates.

Model C: Special Regression Design. Model C is a regression-based evaluation design. Any instrument, rating, or composite measure which correlates highly with the posttest may be used as a pretest. The posttest may be either a normed (Model C1) or non-normed (Model C2) test. If a non-normed test is used, a normed test must also be given to the treatment group sometime during the school year. The model requires that the selection of the treatment group be based exclusively on the pretest measure. All pupils scoring above the cutoff score must be assigned to the comparison group while those scoring below form the treatment group. All pupils in both the treatment and comparison groups must be pre- and posttested.

Post-on-pretest regression lines are calculated separately for the treatment and comparison groups. The treatment group's regression line represents the observed mean posttest performance corresponding to various pretest scores. The comparison group's regression line, when projected across the cutoff score, provides no-treatment estimates for the same pretest scores (illustrations are presented in Chapter VI).

The treatment effect is defined as the distance between the regression lines and is measured separately at two points: at the treatment group's mean pretest score and at the cutoff score. For both measures it is assumed that the project had a positive impact if the observed score is higher than the expected score. However, a substantial difference between the two measures may signal a spurious apparent gain resulting from test ceiling or floor effects.

There are few constraints concerning test selection. Model C does not require the same test for pretest and posttest or that the tests be administered at norming times. It is essential, however, that the pretest and posttest scores be highly correlated and that the instruments not be so difficult or so easy that test floors or ceilings are encountered. The model assumes a linear relationship between the pretest measure and posttest scores; where floor and ceiling effects exist, the pretest/posttest relationship will no longer be linear. Finally, in Model C, scores on the pretest measure must be used as the sole basis for selecting the Title I project participants.

Normal Curve Equivalents

Normal Curve Equivalents (NCEs) are normalized standard scores with a mean of 50 and a standard deviation of 21.06. Because the scale is normalized, it is assumed to be equal-interval--that is, the length of the interval between any two adjacent scores on the scale is equal to the interval between every other adjacent pair of scores.

The fact that the scale is standardized means simply that the scores have been linearly transformed to give them a desired mean and standard deviation. The values of 50 and 21.06* were chosen to relate NCEs to the percentile scale, thereby lending intrinsic meaning to them through the association. The exact relationship is shown in Figure 1.

The two scales match at values of 1, 50, and 99 but, while the NCE scale points are equally spaced, percentiles are widely spaced at

*The NCE standard deviation of 21.06 was derived by dividing the distance from the mean to the 99th percentile (99 - 50 = 49 percentile points) by the same distance measured in terms of normal curve standard deviation units (2.3267 - 0 = 2.3267). The result (49 ÷ 2.3267 = 21.060) yields a scale which includes exactly 98% of the population between values of 1 and 99.

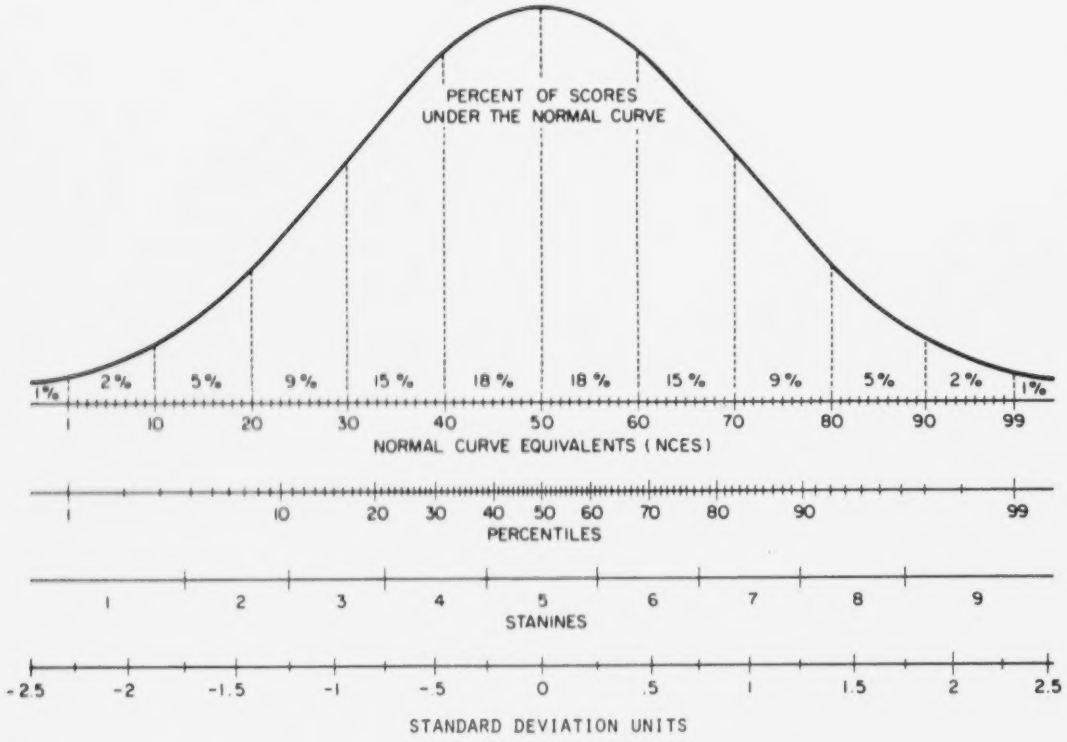

Figure 1. Area under the normal curve divided into NCEs, percentiles, stanines, and standard deviation units.

extremes of the distribution and tightly clustered near the center. It can be seen in Figure 1, for example, that the distance between NCEs of 10 and 20 is the same as the distance between NCEs of 50 and 60. In contrast, the distance between percentiles of 10 and 20 is nearly twice that between percentiles of 50 and 60.

Because percentiles are not all the same size, they should not be used in arithmetic computations. The NCE metric is an equal-interval scale, however. Because of this characteristic, it is legitimate to add, subtract, multiply, and divide NCEs.

Figure 1 also illustrates the relationship between NCEs and stanines. Whereas NCEs divide the distance under the normal curve into 99 equal units, stanines divide the same distance into 9 equal-sized units. Thus, there are 11 NCEs for every one stanine. NCEs closely resemble stanines and both scales are equal-interval, however, NCEs are a more finely grained metric.

Measuring Gains in NCEs with Nationally Normed Tests. From each model, it is possible to derive a no-treatment expectation (via normative data, control group data, or comparison group data) which can be compared against observed post-treatment performance to yield a measure of project impact. When nationally normed tests are used pre and post, observed and expected posttest scores are typically expressed in terms of percentile ranks with respect to a national norm or in terms of the NCE counterparts of these percentiles.

Table 1, which was derived from the table of areas under the normal curve found in most statistics books, gives the NCE that corresponds to each percentile (all numbers are expressed to three significant figures). By using Table 1, observed and expected posttest percentiles can be converted directly to NCEs. The NCE gain is the difference between the observed NCE and the expected NCE.

To illustrate, assume that a norm-referenced evaluation was conducted using a nationally normed test. The treatment group's pretest

Table 1

The Normal Curve Equivalent Corresponding to Each Tenth of a Percentile

%	NCE	%	NCE	%	NCE	%	NCE	%	NCE
1.0	1.0	6.0	17.3	11.0	24.2	16.0	29.1	21.0	33.0
1.1	1.8	6.1	17.4	11.1	24.3	16.1	29.1	21.1	33.1
1.2	2.5	6.2	17.6	11.2	24.4	16.2	29.2	21.2	33.2
1.3	3.1	6.3	17.8	11.3	24.5	16.3	29.3	21.3	33.2
1.4	3.7	6.4	17.9	11.4	24.6	16.4	29.4	21.4	33.3
1.5	4.3	6.5	18.1	11.5	24.7	16.5	29.5	21.5	33.4
1.6	4.8	6.6	18.3	11.6	24.8	16.6	29.6	21.6	33.4
1.7	5.4	6.7	18.4	11.7	24.9	16.7	29.7	21.7	33.5
1.8	5.6	6.8	18.6	11.8	25.0	16.8	29.7	21.8	33.6
1.9	6.3	6.9	18.8	11.9	25.1	16.9	29.8	21.9	33.7
2.0	6.7	7.0	18.9	12.0	25.3	17.0	29.9	22.0	33.7
2.1	7.2	7.1	19.1	12.1	25.4	17.1	30.0	22.1	33.8
2.2	7.6	7.2	19.2	12.2	25.5	17.2	30.1	22.2	33.9
2.3	8.0	7.3	19.4	12.3	25.6	17.3	30.2	22.3	34.0
2.4	8.3	7.4	19.5	12.4	25.7	17.4	30.2	22.4	34.0
2.5	8.7	7.5	19.7	12.5	25.8	17.5	30.3	22.5	34.1
2.6	9.1	7.6	19.8	12.6	25.9	17.6	30.4	22.6	34.2
2.7	9.4	7.7	20.0	12.7	26.0	17.7	30.5	22.7	34.2
2.8	9.7	7.8	20.1	12.8	26.1	17.8	30.6	22.8	34.3
2.9	10.1	7.9	20.3	12.9	26.2	17.9	30.6	22.9	34.4
3.0	10.4	8.0	20.4	13.0	26.3	18.0	30.7	23.0	34.4
3.1	10.7	8.1	20.5	13.1	26.4	18.1	30.8	23.1	34.5
3.2	11.0	8.2	20.7	13.2	26.5	18.2	30.9	23.2	34.6
3.3	11.3	8.3	20.8	13.3	26.6	18.3	31.0	23.3	34.6
3.4	11.6	8.4	21.0	13.4	26.7	18.4	31.0	23.4	34.7
3.5	11.9	8.5	21.1	13.5	26.8	18.5	31.1	23.5	34.8
3.6	12.1	8.6	21.2	13.6	26.9	18.6	31.2	23.6	34.9
3.7	12.4	8.7	21.4	13.7	27.0	18.7	31.3	23.7	34.9
3.8	12.6	8.8	21.5	13.8	27.1	18.8	31.4	23.8	35.0
3.9	12.9	8.9	21.6	13.9	27.1	18.9	31.4	23.9	35.1
4.0	13.1	9.0	21.8	14.0	27.2	19.0	31.5	24.0	35.1
4.1	13.4	9.1	21.9	14.1	27.3	19.1	31.6	24.1	35.2
4.2	13.6	9.2	22.0	14.2	27.4	19.2	31.7	24.2	35.3
4.3	13.8	9.3	22.1	14.3	27.5	19.3	31.7	24.3	35.3
4.4	14.1	9.4	22.3	14.4	27.6	19.4	31.8	24.4	35.4
4.5	14.3	9.5	22.4	14.5	27.7	19.5	31.9	24.5	35.5
4.6	14.5	9.6	22.5	14.6	27.8	19.6	32.0	24.6	35.5
4.7	14.7	9.7	22.6	14.7	27.9	19.7	32.0	24.7	35.6
4.8	14.9	9.8	22.8	14.8	28.0	19.8	32.1	24.8	35.7
4.9	15.1	9.9	22.9	14.9	28.1	19.9	32.2	24.9	35.7
5.0	15.4	10.0	23.0	15.0	28.2	20.0	32.3	25.0	35.8
5.1	15.6	10.1	23.1	15.1	28.3	20.1	32.3	25.1	35.9
5.2	15.8	10.2	23.3	15.2	28.4	20.2	32.4	25.2	35.9
5.3	16.0	10.3	23.4	15.3	28.4	20.3	32.5	25.3	36.0
5.4	16.2	10.4	23.5	15.4	28.5	20.4	32.6	25.4	36.1
5.5	16.3	10.5	23.6	15.5	28.6	20.5	32.6	25.5	36.1
5.6	16.5	10.6	23.7	15.6	28.7	20.6	32.7	25.6	36.2
5.7	16.7	10.7	23.8	15.7	28.8	20.7	32.8	25.7	36.3
5.8	16.9	10.8	23.9	15.8	28.9	20.8	32.9	25.8	36.3
5.9	17.1	10.9	24.1	15.9	29.0	20.9	32.9	25.9	36.4

Table 1 continued

%	NCE	%	NCE	%	NCE	%	NCE	%	NCE
26.0	36.5	31.0	39.6	36.0	42.5	41.0	45.2	46.0	47.9
26.1	36.5	31.1	39.6	36.1	42.5	41.1	45.3	46.1	47.9
26.2	36.6	31.2	39.7	36.2	42.6	41.2	45.3	46.2	48.0
26.3	36.6	31.3	39.7	36.3	42.6	41.3	45.4	46.3	48.0
26.4	36.7	31.4	39.8	36.4	42.7	41.4	45.4	46.4	48.1
26.5	36.8	31.5	39.9	36.5	42.7	41.5	45.5	46.5	48.2
26.6	36.8	31.6	39.9	36.6	42.8	41.6	45.5	46.6	48.2
26.7	36.9	31.7	40.0	36.7	42.8	41.7	45.6	46.7	48.3
26.8	37.0	31.8	40.0	36.8	42.9	41.8	45.6	46.8	48.3
26.9	37.0	31.9	40.1	36.9	43.0	41.9	45.7	46.9	48.4
27.0	37.1	32.0	40.1	37.0	43.0	42.0	45.8	47.0	48.4
27.1	37.2	32.1	40.2	37.1	43.1	42.1	45.8	47.1	48.5
27.2	37.2	32.2	40.3	37.2	43.1	42.2	45.9	47.2	48.5
27.3	37.3	32.3	40.3	37.3	43.2	42.3	45.9	47.3	48.6
27.4	37.3	32.4	40.4	37.4	43.2	42.4	46.0	47.4	48.6
27.5	37.4	32.5	40.4	37.5	43.3	42.5	46.0	47.5	48.7
27.6	37.5	32.6	40.5	37.6	43.3	42.6	46.1	47.6	48.7
27.7	37.5	32.7	40.6	37.7	43.4	42.7	46.1	47.7	48.8
27.8	37.6	32.8	40.6	37.8	43.5	42.8	46.2	47.8	48.8
27.9	37.7	32.9	40.7	37.9	43.5	42.9	46.2	47.9	48.9
28.0	37.7	33.0	40.7	38.0	43.6	43.0	46.3	48.0	48.9
28.1	37.8	33.1	40.8	38.1	43.6	43.1	46.3	48.1	49.0
28.2	37.8	33.2	40.9	38.2	43.7	43.2	46.4	48.2	49.0
28.3	37.9	33.3	40.9	38.3	43.7	43.3	46.4	48.3	49.1
28.4	38.0	33.4	41.0	38.4	43.8	43.4	46.5	48.4	49.2
28.5	38.0	33.5	41.0	38.5	43.9	43.5	46.6	48.5	49.2
28.6	38.1	33.6	41.1	38.6	43.9	43.6	46.6	48.6	49.3
28.7	38.2	33.7	41.1	38.7	44.0	43.7	46.7	48.7	49.3
28.8	38.2	33.8	41.2	38.8	44.0	43.8	46.7	48.8	49.4
28.9	38.3	33.9	41.3	38.9	44.1	43.9	46.8	48.9	49.4
29.0	38.3	34.0	41.3	39.0	44.1	44.0	46.8	49.0	49.5
29.1	38.4	34.1	41.4	39.1	44.2	44.1	46.9	49.1	49.5
29.2	38.5	34.2	41.4	39.2	44.2	44.2	46.9	49.2	49.6
29.3	38.5	34.3	41.5	39.3	44.3	44.3	47.0	49.3	49.6
29.4	38.6	34.4	41.5	39.4	44.3	44.4	47.0	49.4	49.7
29.5	38.7	34.5	41.6	39.5	44.4	44.5	47.1	49.5	49.7
29.6	38.7	34.6	41.7	39.6	44.4	44.6	47.1	49.6	49.8
29.7	38.8	34.7	41.7	39.7	44.5	44.7	47.2	49.7	49.8
29.8	38.8	34.8	41.8	39.8	44.6	44.8	47.2	49.8	49.9
29.9	38.9	34.9	41.8	39.9	44.6	44.9	47.3	49.9	49.9
30.0	39.0	35.0	41.9	40.0	44.7	45.0	47.4	50.0	50.0
30.1	39.0	35.1	41.9	40.1	44.7	45.1	47.4	50.1	50.1
30.2	39.1	35.2	42.0	40.2	44.8	45.2	47.5	50.2	50.1
30.3	39.1	35.3	42.1	40.3	44.8	45.3	47.5	50.3	50.2
30.4	39.2	35.4	42.1	40.4	44.9	45.4	47.6	50.4	50.2
30.5	39.3	35.5	42.2	40.5	44.9	45.5	47.6	50.5	50.3
30.6	39.3	35.6	42.2	40.6	45.0	45.6	47.7	50.6	50.3
30.7	39.4	35.7	42.3	40.7	45.0	45.7	47.7	50.7	50.4
30.8	39.4	35.8	42.3	40.8	45.1	45.8	47.8	50.8	50.4
30.9	39.5	35.9	42.4	40.9	45.2	45.9	47.8	50.9	50.5

Table 1 continued

%	NCE	%	NCE	%	NCE	%	NCE	%	NCE
51.0	50.5	56.0	53.2	61.0	55.9	66.0	58.7	71.0	61.7
51.1	50.6	56.1	53.2	61.1	55.9	66.1	58.7	71.1	61.7
51.2	50.6	56.2	53.3	61.2	56.0	66.2	58.8	71.2	61.8
51.3	50.7	56.3	53.3	61.3	56.0	66.3	58.9	71.3	61.8
51.4	50.7	56.4	53.4	61.4	56.1	66.4	58.9	71.4	61.9
51.5	50.8	56.5	53.4	61.5	56.1	66.5	59.0	71.5	62.0
51.6	50.8	56.6	53.5	61.6	56.2	66.6	59.0	71.6	62.0
51.7	50.9	56.7	53.6	61.7	56.3	66.7	59.1	71.7	62.1
51.8	51.0	56.8	53.6	61.8	56.3	66.8	59.1	71.8	62.2
51.9	51.0	56.9	53.7	61.9	56.4	66.9	59.2	71.9	62.2
52.0	51.1	57.0	53.7	62.0	56.4	67.0	59.3	72.0	62.3
52.1	51.1	57.1	53.8	62.1	56.5	67.1	59.3	72.1	62.3
52.2	51.2	57.2	53.8	62.2	56.5	67.2	59.4	72.2	62.4
52.3	51.2	57.3	53.9	62.3	56.6	67.3	59.4	72.3	62.5
52.4	51.3	57.4	53.9	62.4	56.7	67.4	59.5	72.4	62.5
52.5	51.3	57.5	54.0	62.5	56.7	67.5	59.6	72.5	62.6
52.6	51.4	57.6	54.0	62.6	56.8	67.6	59.6	72.6	62.7
52.7	51.4	57.7	54.1	62.7	56.8	67.7	59.7	72.7	62.7
52.8	51.5	57.8	54.1	62.8	56.9	67.8	59.7	72.8	62.8
52.9	51.5	57.9	54.2	62.9	56.9	67.9	59.8	72.9	62.8
53.0	51.6	58.0	54.2	63.0	57.0	68.0	59.9	73.0	62.9
53.1	51.6	58.1	54.3	63.1	57.0	68.1	59.9	73.1	63.0
53.2	51.7	58.2	54.4	63.2	57.1	68.2	60.0	73.2	63.0
53.3	51.7	58.3	54.4	63.3	57.2	68.3	60.0	73.3	63.1
53.4	51.8	58.4	54.5	63.4	57.2	68.4	60.1	73.4	63.2
53.5	51.8	58.5	54.5	63.5	57.3	68.5	60.1	73.5	63.2
53.6	51.9	58.6	54.6	63.6	57.3	68.6	60.2	73.6	63.3
53.7	52.0	58.7	54.6	63.7	57.4	68.7	60.3	73.7	63.4
53.8	52.0	58.8	54.7	63.8	57.4	68.8	60.3	73.8	63.4
53.9	52.1	58.9	54.7	63.9	57.5	68.9	60.4	73.9	63.5
54.0	52.1	59.0	54.8	64.0	57.5	69.0	60.4	74.0	63.5
54.1	52.2	59.1	54.8	64.1	57.6	69.1	60.5	74.1	63.6
54.2	52.2	59.2	54.9	64.2	57.7	69.2	60.6	74.2	63.7
54.3	52.3	59.3	55.0	64.3	57.7	69.3	60.6	74.3	63.7
54.4	52.3	59.4	55.0	64.4	57.8	69.4	60.7	74.4	63.8
54.5	52.4	59.5	55.1	64.5	57.8	69.5	60.7	74.5	63.9
54.6	52.4	59.6	55.1	64.6	57.9	69.6	60.8	74.6	63.9
54.7	52.5	59.7	55.2	64.7	57.9	69.7	60.9	74.7	64.0
54.8	52.5	59.8	55.2	64.8	58.0	69.8	60.9	74.8	64.1
54.9	52.6	59.9	55.3	64.9	58.1	69.9	61.0	74.9	64.1
55.0	52.6	60.0	55.3	65.0	58.1	70.0	61.0	75.0	64.2
55.1	52.7	60.1	55.4	65.1	58.2	70.1	61.1	75.1	64.3
55.2	52.8	60.2	55.4	65.2	58.2	70.2	61.2	75.2	64.3
55.3	52.8	60.3	55.5	65.3	58.3	70.3	61.2	75.3	64.4
55.4	52.9	60.4	55.6	65.4	58.3	70.4	61.3	75.4	64.5
55.5	52.9	60.5	55.6	65.5	58.4	70.5	61.3	75.5	64.5
55.6	53.0	60.6	55.7	65.6	58.5	70.6	61.4	75.6	64.6
55.7	53.0	60.7	55.7	65.7	58.5	70.7	61.5	75.7	64.7
55.8	53.1	60.8	55.8	65.8	58.6	70.8	61.5	75.8	64.7
55.9	53.1	60.9	55.8	65.9	58.6	70.9	61.6	75.9	64.8

Table 1 continued

%	NCE	%	NCE	%	NCE	%	NCE	%	NCE
76.0	64.9	81.0	68.5	86.0	72.8	91.0	78.2	96.0	86.9
76.1	64.9	81.1	68.6	86.1	72.9	91.1	78.4	96.1	87.1
76.2	65.0	81.2	68.6	86.2	72.9	91.2	78.5	96.2	87.4
76.3	65.1	81.3	68.7	86.3	73.0	91.3	78.6	96.3	87.6
76.4	65.1	81.4	68.8	86.4	73.1	91.4	78.8	96.4	87.9
76.5	65.2	81.5	68.9	86.5	73.2	91.5	78.9	96.5	88.1
76.6	65.3	81.6	69.0	86.6	73.3	91.6	79.0	96.6	88.4
76.7	65.4	81.7	69.0	86.7	73.4	91.7	79.2	96.7	88.7
76.8	65.4	81.8	69.1	86.8	73.5	91.8	79.3	96.8	89.0
76.9	65.5	81.9	69.2	86.9	73.6	91.9	79.5	96.9	89.3
77.0	65.6	82.0	69.3	87.0	73.7	92.0	79.6	97.0	89.6
77.1	65.6	82.1	69.4	87.1	73.8	92.1	79.7	97.1	89.9
77.2	65.7	82.2	69.4	87.2	73.9	92.2	79.9	97.2	90.3
77.3	65.8	82.3	69.5	87.3	74.0	92.3	80.0	97.3	90.6
77.4	65.8	82.4	69.6	87.4	74.1	92.4	80.2	97.4	90.9
77.5	65.9	82.5	69.7	87.5	74.2	92.5	80.3	97.5	91.3
77.6	66.0	82.6	69.8	87.6	74.3	92.6	80.5	97.6	91.7
77.7	66.0	82.7	69.8	87.7	74.4	92.7	80.6	97.7	92.0
77.8	66.1	82.8	69.9	87.8	74.5	92.8	80.8	97.8	92.4
77.9	66.2	82.9	70.0	87.9	74.6	92.9	80.9	97.9	92.8
78.0	66.3	83.0	70.1	88.0	74.7	93.0	81.1	98.0	93.3
78.1	66.3	83.1	70.2	88.1	74.9	93.1	81.2	98.1	93.7
78.2	66.4	83.2	70.3	88.2	75.0	93.2	81.4	98.2	94.4
78.3	66.5	83.3	70.3	88.3	75.1	93.3	81.6	98.3	94.6
78.4	66.6	83.4	70.4	88.4	75.2	93.4	81.7	98.4	95.2
78.5	66.6	83.5	70.5	88.5	75.3	93.5	81.9	98.5	95.7
78.6	66.7	83.6	70.6	88.6	75.4	93.6	82.1	98.6	96.3
78.7	66.8	83.7	70.7	88.7	75.5	93.7	82.2	98.7	96.9
78.8	66.8	83.8	70.8	88.8	75.6	93.8	82.4	98.8	97.5
78.9	66.9	83.9	70.9	88.9	75.7	93.9	82.6	98.9	98.2
79.0	67.0	84.0	70.9	89.0	75.8	94.0	82.7	99.0	99.0
79.1	67.1	84.1	71.0	89.1	75.9	94.1	82.9		
79.2	67.1	84.2	71.1	89.2	76.1	94.2	83.1		
79.3	67.2	84.3	71.2	89.3	76.2	94.3	83.3		
79.4	67.3	84.4	71.3	89.4	76.3	94.4	83.5		
79.5	67.4	84.5	71.4	89.5	76.4	94.5	83.7		
79.6	67.4	84.6	71.5	89.6	76.5	94.6	83.8		
79.7	67.5	84.7	71.6	89.7	76.6	94.7	84.0		
79.8	67.6	84.8	71.6	89.8	76.7	94.8	84.2		
79.9	67.7	84.9	71.7	89.9	76.9	94.9	84.4		
80.0	67.7	85.0	71.8	90.0	77.0	95.0	84.6		
80.1	67.8	85.1	71.9	90.1	77.1	95.1	84.9		
80.2	67.9	85.2	72.0	90.2	77.2	95.2	85.1		
80.3	68.0	85.3	72.1	90.3	77.4	95.3	85.3		
80.4	68.0	85.4	72.2	90.4	77.5	95.4	85.5		
80.5	68.1	85.5	72.3	90.5	77.6	95.5	85.7		
80.6	68.2	85.6	72.4	90.6	77.7	95.6	85.9		
80.7	68.3	85.7	72.5	90.7	77.9	95.7	86.2		
80.8	68.3	85.8	72.6	90.8	78.0	95.8	86.4		
80.9	68.4	85.9	72.7	90.9	78.1	95.9	86.6		

score (which constitutes the expected no-treatment posttest performance) was found to be at the 21st percentile while the observed posttest performance was at the 25th percentile. Using Table 1, the NCEs that correspond to percentiles of 21.0 and 25.0 are found to be 33.0 and 35.8. The NCE gain then is 2.8 (the difference between 35.8 and 33.0).

Measuring Gains in NCEs with Non-Normed Tests. Calculating an NCE gain when non-normed tests have been administered is a slightly more complex procedure than when normed tests are used. Without nationally representative normative data, it is generally not possible to obtain national percentile equivalents for the observed and expected posttest scores (Model A2 is an exception to this general rule). In the absence of percentiles, the measures of observed and expected posttest status cannot be translated into NCEs. It is possible, however, to express the gain in NCEs. The formula is

$$\text{T.E.} = \frac{21.06\ (\bar{Y}_t - \hat{\bar{Y}}_t)\ s_n}{\sigma_n\ s_{nn}} \qquad (2)$$

where

$\bar{Y}_t$ = observed mean posttest score of treatment group pupils

$\hat{\bar{Y}}_t$ = expected mean posttest score of treatment group pupils under no-treatment conditions

s_n = standard deviation of treatment group scores on the normed test

s_{nn} = standard deviation of treatment group scores on the non-normed test

σ_n = standard deviation of norm group scores on the normed test.

The derivation of this formula is presented below. It is not required for an understanding of the system, however, and is presented solely to satisfy the curiosity of interested readers. Others should skip ahead to Measuring Gains in NCEs Using Tests with Local Norms.

Equation (1) from page 4 can be written

$$T.E. = \bar{Y}_t - \hat{\bar{Y}}_t \qquad (3)$$

where

$\bar{Y}_t$ = observed mean posttest score of treatment group pupils

$\hat{\bar{Y}}_t$ = expected mean posttest score of treatment group pupils under no-treatment conditions.

To express Equation (3) in NCEs, both $\bar{Y}_t$ and $\hat{\bar{Y}}_t$ must be expressed as deviations from the (estimated) national mean, measured in (estimated) standard deviation units, and linearly transformed to have a mean of 50 and a standard deviation of 21.06

$$T.E. = \left[21.06\left(\frac{\bar{Y}_t - \hat{\bar{Y}}_n}{\hat{\sigma}_{nn}}\right) + 50\right] - \left[21.06\left(\frac{\hat{\bar{Y}}_t - \hat{\bar{Y}}_n}{\hat{\sigma}_{nn}}\right) + 50\right], \qquad (4)$$

where

$\hat{\bar{Y}}_n$ = estimated mean posttest score of a nationally representative sample

$\hat{\sigma}_{nn}$ = estimated posttest standard deviation of a nationally representative sample on the non-norm test

Equation (4) can be simplified to read

$$T.E. = 21.06\left[\frac{\bar{Y}_t - \hat{\bar{Y}}_n}{\hat{\sigma}_{nn}} - \frac{\hat{\bar{Y}}_t - \hat{\bar{Y}}_n}{\hat{\sigma}_{nn}}\right]. \qquad (5)$$

Equation (5) can be further simplified to read

$$T.E. = 21.06 \left[\frac{\bar{Y}_t - \hat{\bar{Y}}_t}{\hat{\sigma}_{nn}} \right] . \qquad (6)$$

The standard deviation of the national sample is estimated from scores on the nationally normed test which must also be administered under the assumption that the ratio of the local to the national standard deviation will be the same for the normed test as for the non-normed test. In other words

$$\frac{s_n}{\sigma_n} = \frac{s_{nn}}{\hat{\sigma}_{nn}} \qquad (7)$$

where

s_n = standard deviation of treatment group scores on the normed test

s_{nn} = standard deviation of treatment group scores on the non-normed test

σ_n = standard deviation of norm group scores on the normed test.

Solving for $\hat{\sigma}_{nn}$ the equation becomes

$$\hat{\sigma}_{nn} = \frac{\sigma_n \, s_{nn}}{s_n} . \qquad (8)$$

Substituting Equation (8) into Equation (6), we have

$$T.E. = \frac{21.06\ (\bar{Y}_t - \hat{\bar{Y}}_t)\ s_n}{\sigma_n\ s_{nn}} \quad . \tag{9}$$

All of the data required to solve Equation (9) [which is identical to Equation (2)] should be in hand if Model B2 or C2 has been correctly implemented. $\bar{Y}_t$, $\hat{\bar{Y}}_t$, s_n, and s_{nn} are all calculated from treatment or comparison group test scores and σ_n is obtained from the test publisher's manual.

Measuring Gains in NCEs Using Tests with Local Norms. When tests are used that have only local and no national norms, the gain is calculated using the procedure employed with non-normed tests. A nationally normed test must be given in order to derive an estimate of a national sample's standard deviation on the locally normed test. The necessary values are inserted in Equation (2) and the NCE gain can then be calculated.

When tests are given which have both local and national norms, Equation (6) is used instead. The process is relatively straightforward. After the standard score difference between observed and expected posttest performance is determined, it is then divided by the standard deviation of the nationally representative sample's test scores. In contrast to the situations in which non-normed tests are used, the national sample's standard deviation does not have to be estimated; instead it can be obtained from the technical manual for the test. After the standard score gain is divided by the national standard deviation, it is multiplied by 21.06 to obtain the NCE gain.

II. CHOOSING AN EVALUATION MODEL

Models A, B, and C differ from one another in terms of technical credibility and ease of implementation. Unfortunately, the best design from a scientific point of view may be the most difficult to implement in the context of Title I. Thus, practical concerns may necessitate compromising scientific rigor.

To assist evaluation personnel in selecting the most suitable model for their particular situations, this chapter contains a five-step decision tree designed to focus attention on the most relevant factors in the decision-making process. The decision tree, illustrated in Figure 2, presents the evaluation models in a hierarchy based on technical desirability. Since choosing a model will always involve making tradeoffs between technical and practical considerations, an attempt is made to enumerate the latter and to assist the evaluator in making the best possible decision.

Model B, the control group model, is the most scientifically rigorous of the three evaluation designs. Its technical merit, however, depends on the similarity between treatment and control groups. Ideally, pupils drawn from a single population should be randomly assigned to the two groups. The model is slightly less credible when the groups are not formed through random assignment but appear to be highly similar in all educationally relevant ways. Even where there are distinct differences between treatment and control groups, statistical adjustments can be made (as long as the differences are not too large) so that Model B remains the preferred choice over Models A and C.

Model C, the special regression model, is the clear second choice over Model A, the norm-referenced model, since it, like Model B, quantifies and compensates for initial differences between groups. In Model A, it can be assumed that children in the treatment group are

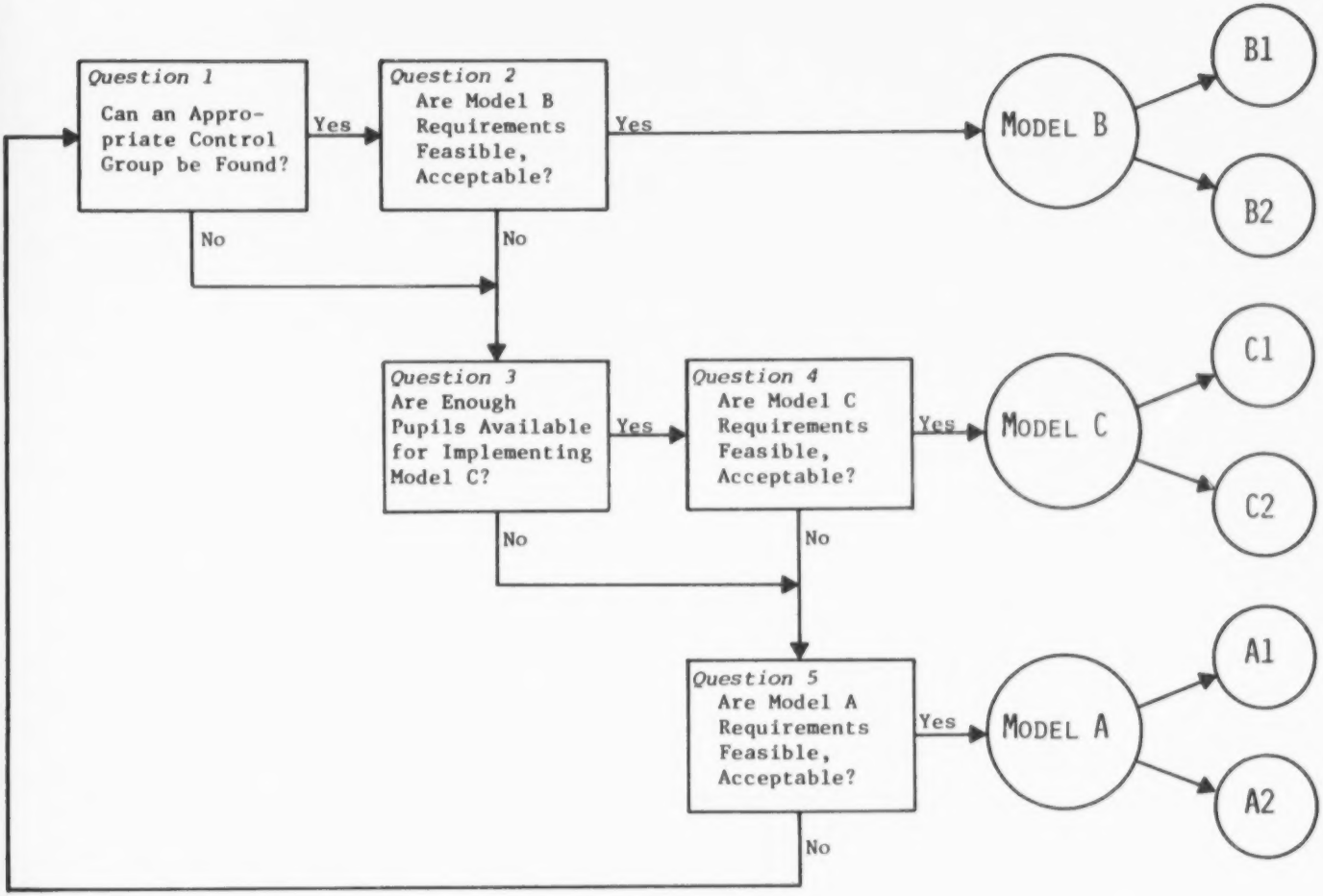

Figure 2. Decision tree for selecting an evaluation model.

not exactly like those in the comparison (norm) group. It is not possible, however, either to determine the nature of the differences or to compensate for them.

Despite the clear order of preference, all three of the models should yield valid evaluation results if properly implemented. Feasibility and other practical considerations may well counterbalance scientific rigor, and it is certainly the case that a well implemented Model A will yield more credible results than a poorly implemented Model B or C.

The remainder of this chapter is tied directly to the questions composing the decision tree. While there are only five of them, each has complex implications. In some instances, the answers to the questions will be immediately apparent. In other cases, careful thought will be required to weigh all the possible outcomes and make the best choice among the alternatives. While every attempt has been made to provide clear-cut decision criteria in the discussion following each question, it has not always been possible to do so, and considered judgments based on complex technical issues will often be required.

Before making a final choice among the models, it would perhaps be wise to review and re-examine the decisions made initially and to work through the decision tree a second time.

Question 1

Can a suitable control group be found?

Yes. Proceed to Question 2.

No. Skip to Question 3.

Discussion

Ideally, treatment and control groups should be identical. The pupils in the two groups should match one another in terms of age, sex, intelligence, race, ethnicity, socioeconomic status, pretest achievement levels, and all other educationally relevant variables.

Theoretically, this ideal could be attained if pairs of identical pupils could be found and one member of each pair were assigned to each group. In practice, however, this approach borders on the impossible. In fact, it is often not clear exactly what pupil characteristics are educationally relevant or to what extent individual pupils possess them.

In most experimental settings, assignment to groups is based on some random event such as the flip of a coin. While this approach does not produce identical groups, it tends to minimize differences. Probability theory, furthermore, enables us to determine precisely how frequently differences of any specified size will occur by chance. Finally, statistical procedures have been developed which adjust the posttest scores of the treatment and control groups for any differences in pretest scores resulting from random assignment. The set of procedures which accomplishes this adjustment is called covariance analysis.

In Title I contexts, it is generally not feasible to establish treatment and control groups through random assignment processes since the legislation requires that the program serve those children with

the greatest educational need.* It sometimes happens, however, that children in a noneligible school in the district (or in an adjacent district) resemble the Title I children enough to provide a useful baseline against which to assess growth. If such groups can be found, Model B should be employed since it is superior in most respects to both Model C and Model A.

In some situations it will be clear that Model B is not feasible because no group even remotely similar to the Title I students can be found to serve as a control. In other situations, similar groups will be available, but there will be some question as to whether or not they sufficiently resemble the treatment group to provide a valid baseline for comparison. The following paragraphs are addressed to this question.

In any practical application of Model B, it is not necessary that the treatment and control groups be identical. It is not necessary that they be formed by random assignment or even that they be as much alike as they would have been had they been formed by random assignment. They must only be enough alike so that the accuracy of Model B's no-treatment expectation equals or exceeds that of Models C or A.

The accuracy of the no-treatment expectation in Model C depends on the number of children in the comparison group, the size of the correlation between their pre- and posttest scores, and the range of their pretest scores. A large number of pupils and a high correlation would be required before we could begin to have as much confidence in Model C as we would in Model B implemented with random assignment of pupils

*An exception to this generalization would exist in cases where the Title I treatment lasted half a year or less. Under these circumstances children could be randomly assigned to first- and second-semester (for example) groups. Then during the first semester, the second semester children could serve as a true control group.

to treatment and control groups. Where the groups are not randomly assigned, however, the credibility of Model B decreases as the difference between groups increases. The point at which Model C becomes preferable to Model B, however, cannot be precisely defined except where there is random assignment.

The accuracy of Model A's no-treatment expectation depends on the similarity between the treatment group children and those children in the normative sample who obtained the same test scores. (Note: the norm-referenced model does not compare treatment group children against "average" children or against "the national norm.") If it can be assumed that children in the treatment group are more like those in the norming sample than like children in a candidate control school, then Model A should be preferred over Model B. If, on the other hand, it is clear that the treatment and candidate control groups are more similar than the treatment and norm groups (as they certainly would be, for example, if the student population in both schools were largely Native American), then Model B should be preferred over Model A. (It should be noted that Model A and Model B can be implemented simultaneously with only a few additional procedural constraints beyond those required for implementing Model B alone.)

There is, unfortunately, no way of determining the relevant characteristics of those children in national norming samples who obtained the same scores as the treatment group students. It is clear, however, that they will not all be urban or rural or black or poor. Thus, if the treatment group is homogeneous in one of these respects (or another respect of potential relevance), extra effort should be made to find a suitable control group, to adopt Model C, or to develop local norms.

Model A should be regarded as a special case of the control group design where children in the norming sample serve, in essence, as controls. The question should then be asked, "Can a better control group

be found than one composed of pupils in the norming sample who obtained the same scores as the treatment group pupils?"

Clearly no hard and fast rules can be provided regarding the adequacy of control groups. Good judgment must be exercised in deciding whether or not an available group is sufficiently similar to the treatment group to serve as a control.

One rule of thumb can be offered. If, on the pretest, the treatment and control groups are found to differ by more than four (4) NCEs (or by a fifth of a standard deviation with respect to the national norm), then use of Model C or Model A might be preferable. Even this rule of thumb, however, should be relaxed when the conditions for implementing one of the other models are also less than ideal.

One final point must be made. It is absolutely essential that the pupils in the control group be selected in exactly the same manner as the pupils in the treatment group. The often-encountered practice of selecting low-scoring pupils out of superior groups to match the achievement levels of pupils in Title I groups invalidates Model B. This issue is discussed further in the chapter of this guidebook which covers implementation procedures for Model B, and in the pamphlet entitled The Regression Effect.

Question 2

Can the requirements imposed by Model B be met, and are they acceptable?

Yes. Proceed to Model B Implementation Instructions.

No. Go on to Question 3.

Discussion

There are only three operational requirements imposed by Model B that may be difficult to implement:

a. The treatment and control groups must be pretested and posttested at the same time.

b. All supplementary instructional services must be withheld from the control group during the entire interval between pretesting and posttesting.

c. The children must be selected for the treatment and control groups in the same manner.

These issues are discussed separately below.

If at all possible, treatment and control groups should be pretested on the same day and at the same time during that day. The same schedule and procedure should be followed at posttest time. If this schedule cannot be met, the treatment and control groups should be tested as close to the same time as possible. Certainly no more than a few days should elapse between the times the two groups are tested. Furthermore, the same testing-time delay (if any exists) should be maintained for both pretesting and posttesting. In other words, if the treatment group is pretested Monday morning and the control group Wednesday afternoon, posttesting should be staggered correspondingly.

All testing should be done by persons who are thoroughly trained in the proper test-administration procedures and who are disinterested in the outcomes of the evaluation.

The control group is used to generate a no-treatment expectation for the Title I pupils. This means that its exposure to instruction during the time interval between pretest and posttest should resemble as closely as possible the exposure the treatment group would have experienced had it not participated in the Title I treatment (or any substitute for it). Clearly, if the control group students participate in a State-supported compensatory education project, their posttest scores do not provide a reasonable no-treatment expectation. Under these circumstances, the Model B comparison would not address the question of whether or not the Title I treatment was effective. Instead, it would determine which treatment was the more effective, State or Federal.

The need to select treatment and control group members in exactly the same manner can present significant technical problems, and the reader is cautioned to study the appropriate portions of the Model B implementation procedures carefully.

Model B may be implemented using either normed or non-normed (criterion-referenced) tests. If a non-normed test is used (Model B2), a test with national norms must also be administered. The normed test is used solely to convert the measured gain into NCE units.

If Model B is implemented with a normed test, it would be desirable, although certainly not necessary, to test at times corresponding to empirical normative data points. Following this practice enables the simultaneous implementation of Model A.

Question 3

Are there enough participants and nonparticipants at each grade level to enable implementation of Model C?

Yes. Proceed to Question 4.

No. Skip to Question 5.

Discussion

The accuracy of Model C's no-treatment expectation is a function of three conditions:

a. The stability of the treatment group's mean pretest score (a function of the number of cases in the treatment group).

b. The stability of the comparison group's post-on-pretest regression line (a function of the number of cases in the comparison group and the correlation between their pre- and posttest scores).

c. The distance that the comparison group's regression line must be projected to meet the treatment group's mean pretest score (a function of the proportions of pupils in the treatment and comparison groups).

Of these conditions, the stability of the regression line is by far the most important. Its primary determiner, furthermore, is the number of pupils in the comparison group. Unfortunately, while sample size is the most important consideration in assessing the utility of Model C, it is not possible to specify exactly how large it must be without information about the pretest-posttest correlation and the proportions of pupils in the two groups.

If Model C is properly implemented (i.e., test ceiling and floor effects are avoided, the integrity of the cutoff score is maintained,

etc.), the error associated with its no-treatment expectation is random error. Thus, in any given case, the expected value of the error is zero. The same expectation applies to Model B implemented with random assignment. Model A and Model B when implemented without random assignment, on the other hand, involve systematic as well as random error components. It is also the case, unfortunately, that the size of the systematic error component in these models cannot be precisely estimated. Thus, while Models A and B entail smaller amounts of random error than Model C, Model C should produce no systematic error; it should thus be preferred to both Models A and B where very large sample sizes are available and where treatment and comparison (or normative) groups are clearly dissimilar.

When choosing between Model C and Model A, good judgment will be required because no clear-cut decision criteria can be specified. Generally Model C will be preferable if large numbers of treatment and comparison group chidren are available (ideally, about 50 treatment and 100 comparison pupils), and a high correlation can reasonably be expected between pre- and posttest scores.*

*Adequately high correlations can generally be expected with children at the third grade and above if test reliabilities are high. Test ceiling and floor effects must, however, be carefully avoided.

Question 4

Can the requirements imposed by Model C be met, and are they acceptable?

Yes. Proceed to Model C Implementation Instructions

No. Proceed to Question 5.

Discussion

There are only three operational requirements imposed by Model C that may be difficult to implement:

a. Pretesting is required for the entire group from which the treatment and comparison subgroups will subsequently be formed.

b. Assignment to the treatment and comparison subgroups must be based entirely on the pretest measure (although the pretest measure can be a composite of test scores and teacher judgments made independently). All pupils scoring above a selected pretest value must be assigned to the comparison group, while all pupils scoring below this cutoff score must be assigned to the treatment group.

c. All supplementary instructional services must be withheld from the comparison group children throughout the entire period between pretesting and posttesting.

These issues are discussed separately below.

If at all possible, all pupils who will form the treatment and comparison groups should be pretested on the same day and at the same time during that day. The same schedule and procedure should be followed at posttest time. If this schedule cannot be followed, subgroups should be tested as close to the same time as possible. Certainly no more than

a few days should elapse between the times the first and last subgroups are tested. Furthermore, the same testing delay (if any) should be maintained for both pretesting and posttesting. In other words, if one third-grade classroom (for example) is pretested Monday morning, a second classroom on Tuesday morning, and a third classroom on Wednesday afternoon, posttesting of each group should be correspondingly staggered.

All testing should be done by persons who have been thoroughly trained in the proper test administration procedures and who are disinterested in the outcomes of the evaluation.

Assignment of pupils to the treatment and comparison groups must be accomplished solely on the basis of the pretest measure. All pupils falling below a selected cutoff value must be assigned to the treatment group. No pupils scoring above the selected value may be assigned to the treatment group. If it is absolutely essential to make exceptions to this rule, they must be kept to a minimum and all pupils who are exceptions to the rule must be excluded from the evaluation.

The cutoff score is arbitrary. It can be selected after pretesting by counting up from the lowest scoring pupil until the number of pupils which can be served with available funds is reached. The cutoff score should then be drawn halfway between the highest score of a treatment group child and the lowest score of a comparison group child.

The pretest measure may be a composite of a test score and an independently made teacher judgment. It may even be the composite of several test scores and several judgments. Any judgments must be made independently (without knowledge) of one another, however, and independently of the test scores. It is not acceptable to allow teachers to move specific children into or out of the treatment group because their test scores are judged to be too high or too low (see Technical Paper No. 10 entitled Composite Scores.)

The comparison group data are used to generate a no-treatment expectation for the Title I pupils. This implies that the comparison

group's exposure to instruction during the time interval between pretest and posttest should resemble as closely as possible the exposure the treatment group would have had if it had not participated in the Title I treatment (or a substitute for it). If any portion of the comparison group participates in a State-supported compensatory education project, their scores cannot be used to generate a no-treatment expectation. The treatment, or substitutes for it, must be withheld from all pupils scoring above the cutoff point.

It is necessary in Model C to obtain both pretest and posttest scores on pupils in both the treatment and comparison groups. Ideally, all pupils in both groups should be both pretested and posttested, although, with large enough numbers, some sampling may be permitted at posttest time only. Posttesting should be accomplished in accordance with the same guidelines outlined above for pretesting.

Question 5

Can the requirements imposed by Model A be met, and are they acceptable?

Yes. Proceed to Model A Implementation Instructions.

No. You have rejected all available choices. Go back to question 1 and reconsider.

Discussion

There are only three operational requirements imposed by Model A which may be difficult to implement:

a. Pretesting must be accomplished after project participants have been selected. This implies that, if selection is based on test scores, two separate testings must be undertaken (although last year's posttest scores could be used to select this year's participants).

b. Both pretesting and posttesting must be accomplished at times which are close to empirical normative data points for the tests used. If the evaluation is based on non-normed tests (e.g., criterion-referenced tests), they must be administered at times which are close to the norming dates for the standardized test that is used in conjunction with the non-normed tests.

c. The same level and form of the same test should be used both pre and post.

Each of these issues is discussed separately in the following paragraphs.

The problem with using a single test administration for the dual purpose of selecting project participants and assessing their pretreatment achievement status stems from the fact that test scores are not

perfectly reliable indices of ability. They reflect substantial amounts of measurement error. Because the measurement error is necessarily correlated with observed scores (see Technical Paper No. 3 entitled The Regression Effect), the majority of scores below the mean of the group tested contain negative error components while the majority of those above the mean contain positive error components. Put another way, the majority of scores below the mean tend to be spuriously low while the majority of those above the mean tend to be spuriously high. Furthermore, the magnitude of the error influence increases with distance from the mean.

If a group of students, selected because of their low scores, is retested with the same or an alternate form of the test, the majority of them will tend to obtain higher scores on the second testing. This effect occurs because measurement error is randomly distributed within the group tested. Most of the scores above the mean on the second testing will contain positive error while the same student's scores on the first testing were below the mean of the total group tested and thus usually contained negative error. This shift from primarily negative to primarily positive error for half of the selected subgroup is called "regression toward the mean" or "statistical regression" or simply "the regression effect."

If the second testing is the posttest, the gain due to the regression effect will be added to whatever gain resulted from the instructional treatment, and the treatment will appear more effective than it actually was. To remove the biasing influence, the selected group must be retested at the beginning of the treatment period, and the second set of scores must be used as the pretest measure.

Both pretesting and posttesting must be accomplished at times which are close to empirical norming dates. This requirement exists because the norms are being used as a substitute for a control group. Although

the testing-time issue is not quite as critical in norm-referenced evaluations as those employing a control group, the same principal applies. It is clearly inappropriate to compare groups which were tested at significantly different times.

The compilation of normative data is generally accomplished over a period of several weeks. The midpoint of the norming interval should be considered the norming date. The treatment group may be tested at any time within two weeks on either side of the norming date without interpolating the norms, or within six weeks of the norming date using linear interpolation between the surrounding normative data points. Deviations from the norming point should be in the same direction and of approximately the same magnitude for both pretesting and posttesting. In other words, there is no problem if both the pretest and posttest are administered two weeks prior to the corresponding norming dates. It is not appropriate, however, to pretest two weeks early and posttest six weeks early (or two weeks late), even if linear interpolations are made.

While it is not essential to use the same form and level of an achievement test both pre and post, this practice is highly recommended. While most of the commonly used achievement tests provide alternate forms, they are not always highly comparable in terms of either content or difficulty. While there may be slight advantages to using alternate forms where the test publisher has done very careful matching of them on both the content and difficulty dimensions, as a general rule of thumb, it is safer to use a single form for both pretesting and posttesting.

Content differences obviously exist between different levels of tests. While most publishers have done a good job of overlapping the content of adjacent levels and developing a single score scale which spans all levels, the "inter-level articulation" is always less than perfect. All other things being equal, it is generally preferable to stick with the same level for both pretesting and posttesting.

The best justification for changing levels between pretesting and posttesting is to avoid ceiling and/or floor effects. Another problem which frequently arises is that using the same level of a test both pre and post will require pretesting in level and posttesting out of level. It is not clear that this practice is markedly superior to that of changing levels. Furthermore, with some instruments, out-of-level testing is not feasible in norm-referenced evaluations because the publisher has not provided the kind of across-level scale which enables entering the in-level norms tables.

III. IMPLEMENTATION OF MODEL A1

General Characteristics

In contrast to Models B and C, the norm-referenced model requires the fewest and simplest calculations and is probably the least expensive to implement. However, it is also the least rigorous of the three models.

Model A1 requires that a normed achievement test be administered to the treatment-group children as both a pretest and a posttest. All testing must be done within a few weeks of the empirical normative testing dates. The pretest status of the treatment group is determined relative to a nationally or locally representative norm group of children at the same grade level. The no-treatment expectation is that the treatment group pupils would have maintained their percentile status with respect to the norm group from pre- to posttest if they had not received any special treatment. If their posttest status is found to exceed their pretest status, the assumption is made that the improvement resulted from participation in the special project.

Deriving the No-Treatment Expectation

Tests with National Norms. If the evaluation is done using national norms, the no-treatment expectation in NCEs may be derived in several ways. The following method is suitable when testing is done either in-level or out-of-level as long as the standard score scale spans all test levels (i.e., is an expanded standard score scale). If the test does not have an expanded standard score scale, out-of-level testing is not a viable option in Model A1.

Step 1. Convert each pupil's pretest raw score to a standard score.

Step 2. Average the standard scores.

Step 3. Convert the average pretest standard acore to its (in-level) percentile equivalent. [Note: if the test manual provides raw-score-to-percentile conversion tables rather than standard-score-to-percentile conversion tables, the average pretest standard score must first be converted to its (in-level) raw score equivalent and the raw score then converted to its percentile equivalent.]

Step 4. Convert the percentile obtained in Step 3 to an NCE. This NCE is the no-treatment posttest expectation.

An alternate procedure may be chosen when testing is done in-level and the test publisher provides raw-score-to-percentile conversion tables rather than standard-score-to-percentile conversion tables.

Step 1. Convert each pupil's pretest raw score to a percentile.

Step 2. Convert each pupil's pretest percentile to an NCE.

Step 3. Average the pretest NCEs for the entire group. This average NCE is the no-treatment posttest expectation.

Some variation on each of these two methods is possible, but the variations will always be computationally more cumbersome than the procedures involved here. If the test publisher provides raw-score-to-NCE conversion tables, of course, the steps involving conversions to and from percentiles can be eliminated.

Tests with Local Norms. If local norms are used for the evaluation and the test instrument has national norms as well, the no-treatment expectation is derived in terms of standard scores. (See Technical Paper No. 7 entitled Local Norms and Appendix G of the building- and LEA-level instructions.) The procedures are as follows:

Step 1. Convert each pupil's pretest raw score to a standard score.

Step 2. Average the standard scores.

Step 3. Convert the average standard score to its local percentile equivalent.

Step 4. Enter the posttest local norm table with the pretest percentile and read out the corresponding standard score. This score is the no-treatment expectation.

If the locally normed test used for the evaluation has no national norms, the same procedures outlined above should be followed to derive the no-treatment expectation. Under these circumstances, however, raw scores must be used in place of standard scores and a nationally normed test must also be administered. The results from the nationally normed test are used to convert the raw score gain to an NCE gain (see next section).

Regardless of the type of norms used, the validity of the model rests on the assumption that a group will maintain its status from pre- to posttest in the absence of any special treatment. It is likely, however, particularly if national norms are used, that the status of some subgroups will increase while that of others will decrease between testings. The meaningfulness of a norm-referenced evaluation is diminished where such patterns exist. Model A is most appropriate where the treatment-group children are highly similar in terms of age, race, ethnic background, socioeconomic status and other educationally relevant variables to the children in the norming sample.

NCE Gains

When tests with national norms are used, the observed post-treatment performance can be expressed in NCEs by following either of the raw score conversion procedures outlined above. The NCE gain is then simply the NCE posttest score minus the NCE no-treatment expectation.

When the evaluation uses local norms, the NCE gain is derived in a slightly different way. Both the observed post-treatment performance and the no-treatment expectation are expressed in standard-score (or raw score) units. To obtain the NCE gain it is necessary first to find the gain by subtracting the no-treatment expectation from the observed score. The gain is then divided by the standard-score (or raw-score) standard deviation of the national norm group (which must be estimated if the test is not nationally normed).* Finally, this quantity is multiplied by 21.06, to convert it to NCEs.

Rules for Implementation

- The model requires using a test with either national or local norms. If the test has only local norms, a test with national norms must also be administered.
- Whenever possible, the same level and form of the test should be administered at pretest and posttest times. In some cases, implementing this recommendation will require that either the pretest or the posttest be out-of-level. Unfortunately, pretesting in-level and posttesting out-of-level (or vice versa) is also an undesirable practice--perhaps about as undesirable as changing the test levels. As long as the test has an expanded standard score scale, either practice may be adopted. In norm-referenced evaluations, out-of-level testing requires the use of tests with an expanded standard score scale.
- That test level should be selected which is appropriate for the functional level of the children (see Technical Paper No. 6, entitled Out-of-Level Testing). Ideally the average child in

*The procedure for estimating the standard deviation of the national population on a test which has not been nationally normed is given in the chapter Implementation of Models B1 and B2, under the subheading NCE Gains.

the project should answer 50 percent of the test questions correctly. If the average project child answers a third or fewer of the items correctly, then the next lower level of the test will be more suitable.

- All testing must be accomplished no earlier than two weeks before and no later than two weeks after the midpoint of the period during which the normative data were collected, unless the norms are linearly interpolated. Refer to Technical Paper No. 5, entitled Characteristics of Eight Commonly Used, Nationally Normed Tests for the appropriate testing dates and norms tables to use for norm-referenced evaluation. With linear interpolation, testing can be done within six weeks of the midpoint. If beginning-of-year normative data were collected in late September/early October and end-of-year normative data in mid-April, then September 30 and April 15 would be taken as the midpoints of the norming intervals. Testing times between September 15 and October 15 and between April 1 and April 30 are acceptable without interpolation. By interpolating between the surrounding normative data points, testing times can be extended from September 8 to October 22 and March 26 to May 7.

  When pre- and posttesting dates deviate from the norming dates, the deviations should be in the same direction and of approximately the same size.

- Participants in the Title I project must not be chosen on the basis of their pretest scores (see Technical Paper No. 3, entitled The Regression Effect). For the purposes of the model, selection may be based on classroom grades, previous test scores, or expert judgment, but not on the students' pretest scores.

Sequence of Tasks

Step 1: Select a norm-referenced achievement test on the basis of

its content and the time or times of year for which it has empirical norms. If pre- and posttesting of the Title I project children is to be done in the spring, for example, then a test must be selected which has empirical norms in spring. Choose the appropriate level and form to be given to each grade.

Step 2: Select the project participants. For the evaluation to be valid, selection must be based on something other than pretest scores.

Step 3: Pretest the participants within two weeks (or six weeks, using interpolation) of either side of the midpoint of the publisher's empirical normative testing period. Make-up tests should be administered not more than two weeks after the testing of the project children. In administering the test, carefully follow the procedures outlined in the publisher's manual.

Step 4: Score the tests and record the pretest scores.

Step 5: After the project, posttest the project participants. In order to measure its effectiveness, at least two-thirds of the project's instruction must have occurred between pre- and posttesting. Posttesting must be conducted within two weeks (or six weeks, using interpolation) of the midpoint of the empirical normative testing period. Deviation from the midpoint should be in the same direction and of the same size as at pretest time. Make-up posttests should be given within two weeks of the original posttest. Administer all tests according to the procedures outlined in the publisher's manual.

If children have been graduated from the project during the school year and have pretest scores, they should be located and posttested. Their scores can be included in the analysis of project impact.

Step 6: Score the tests and record the posttest scores.

Step 7: Determine the effectiveness of the project by carrying out the calculations as specified in the Model A1 instructions.

IV. IMPLEMENTATION OF MODEL A2

General Characteristics

This model is less rigorous than any of the other models including Model A1 and should only be used when it is not possible to have a control or a comparison group, and when the testing instrument for the evaluation has neither national nor local norms.

Model A2 will usually have higher testing costs than Model A1 because a third test must be administered in addition to the pre- and posttest. However, unlike Models B and C only the Title I project children need to be tested, making the costs less than for these models. The calculations are not difficult, but they will require more time than those involved in Model A1. Often it will be necessary to interpolate to find needed values. In addition, this model imposes the greatest restrictions on testing times. All the tests, normed and non-normed, must be given within a few weeks of the empirical normative testing dates for the normed test.

Model A2 is a norm-referenced design in which a non-normed test is administered to a treatment group at pretest and posttest times. A normed test must also be administered at pretest time. The non-normed pretest scores are equated to the normed pretest scores through the use of an abbreviated version of the equipercentile method. By equating the two tests, normative data compiled for the normed instrument can be used to derive a no-treatment expectation for the non-normed test.

The equating procedure begins with listings of the raw scores, their frequencies, and cumulative frequencies for each of the three tests. The scores on the pretests are equated by assuming (in essence) that two raw scores are equivalent if they have the same cumulative frequency. The procedure is essentially the same as that used by most test

publishers to equate alternate test forms, except that the only equivalency which needs to be computed is that for the median posttest score on the non-normed test.

Deriving the No-Treatment Expectation

The percentile status of the treatment group is determined at pretest time by entering the appropriate norms table with their median pretest standard score on the normed test. This percentile value is the no-treatment posttest expectation since it is assumed that the children would have the same status with respect to the national norms at posttest time if they had not participated in the project. As is the case with Model A1, the validity of Model A2 depends on the degree to which this assumption holds for any group of Title I project participants.

NCE Gains

The observed posttest performance is the treatment group's median raw score on the non-normed posttest. Through the equating process, the normed test score equivalent of this median score is found, and its corresponding national percentile rank is determined from the norms tables. The observed performance and the expected no-treatment performance percentiles are converted to NCEs and the difference between these values is the NCE gain resulting from the project.

Rules for Implementation

- Testing with both the normed and the non-normed test should be done at times which correspond to empirical normative data points for the normed test. (Refer to Technical Paper No. 5, entitled Eight Commonly Used, Nationally Normed Tests for appropriate testing dates.) All testing should be done no earlier than two weeks before and no later than two weeks after the

midpoint of the period during which the normative data were collected, unless the norms are linearly interpolated. With linear interpolation, testing can be done within six weeks of the midpoint.

- The two normed and non-normed tests should measure approximately the same abilities (i.e., both tests measure reading comprehension or both measure vocabulary, etc.). Furthermore, it is essential that they be highly correlated. If the correlation is less than .6 for the treatment group children, the normed and non-normed tests cannot be meaningfully equated, and Model A2 should not be used.
- As in Model A1, participants for the Title I project must not be selected on the basis of the normed or non-normed pretest scores (see Technical Paper No. 3 entitled, The Regression Effect).

Implementation Procedures

Step 1: Select a valid and reliable non-normed test.

Step 2: Select a normed test. It should cover the same subject area as the non-normed test. Its empirical normative data point(s) will determine when all testing will be done.

Step 3: Select the project participants. For the model to be valid, selection must be based on something other than the normed or non-normed pretest scores.

Step 4: Administer the normed test to the treatment group within two weeks (or six weeks, using interpolation) of of the midpoint of the publisher's normative testing period. Make-up tests should be given within two weeks of the original testing session. In administering the test, carefully follow the procedures outlined in the publisher's manual.

Step 5: Administer the non-normed pretest after the normed test, but within the two weeks (or six, with interpolation) of the midpoint of the normative testing period for the normed test. Make-up tests should be given within two weeks of the original testing session.

Step 6: Score the pretests and record the scores on the forms.

Step 7: After completion of the project, posttest the project participants. In order to measure the effectiveness of the treatment, at least two-thirds of it must occur between pre- and posttesting. In addition, posttesting must be conducted within two weeks (six weeks, with interpolation) of the midpoint of the empirical normative testing period for the normed test. Make-up tests should be given within two weeks of the original testing. If children have graduated from the project during the school year and they have both a normed and non-normed pretest score, they should be located and posttested. Their scores can be included in the analysis of project impact.

Step 8: Score the posttest and record the scores.

Step 9: Determine the effectiveness of the project by carrying out the calculations specified in the Model A2 instructions. Only those children who have all three test scores--a normed pretest score, a non-normed pretest score, and a non-normed posttest score--should be included in the evaluation.

V. IMPLEMENTATION OF MODELS B1 AND B2

General Characteristics

Model B, if implemented correctly, is the most rigorous of the models because a control group provides the most accurate no-treatment expectation. The cost of testing is about the same as that in Model C, but is approximately double that of Model A. In general the calculations are slightly more complex than Model A but much less complex than Model C. If the mean pretest scores of the treatment and control groups are highly similar, only average posttest scores must be calculated for each group. If pretest differences exist, it is also necessary to calculate standard deviations and, in some instances, correlation coefficients.

Model B involves testing a treatment and control group at pretest and posttest times with a normed or non-normed test. If a non-normed test is used, at some point during the school year, a normed test must also be administered to the treatment group. The standard deviation of the normed-test scores is used to convert the difference between the observed posttest scores and the no-treatment expectation into units common to all the models, Normal Curve Equivalents.

Selecting a Control Group

If the treatment and control groups are composed of intact classrooms, the process of selecting a control group is not particularly hazardous--as long as the two groups are sufficiently similar (see Chapter II, "Choosing an Evaluation Model"). It is often the case, however, that a Title I project will not involve all of the students in a classroom. Under these circumstances, the control group must be selected in exactly the same manner as the treatment group or differential regression effects will invalidate the evaluation (see Technical Paper No. 3, entitled The Regression Effect). Unfortunately, it

will not always be clear what "selected in exactly the same manner" implies.

A clear understanding of the regression effect will be the most valuable aid to the proper selection of control groups. Consider the following two examples:

Example A

In an eligible school, a decision was made to provide Title I reading instruction to all third-grade pupils in the lowest quartile of the national distribution. Under this criterion, approximately 60 percent of the third graders qualified for the treatment group. The most similar non-eligible school in the district was chosen to provide a control group. Within that school, control children were selected who fell into the lowest quartile of the national distribution. Only 40 percent of the third-grade children in the control school met this criterion.

Example B

The situation was exactly the same as in Example A except that the lowest-scoring 60 percent of the students in the control-school third grade were selected to make up the control group. In both Example A and Example B, the selected control groups will have higher true achievement levels than the treatment groups. This fact seems apparent in Example B but not in Example A.

In Example A, the selected control group is farther below the mean of its class than the treatment group is below the mean of its class. Thus, on retesting, it will show a larger regression effect and will achieve a higher score on the second testing than the treatment group, all others things being equal.

In Example B, it is clear that the control group started out with higher scores than the treatment group. Since the treatment

and control groups, on the other hand, had the same test score relationship to the classes from which they were drawn, both, presumably, would show the same amount of regression toward the mean on retesting.

On the surface it might appear that either the Example A or the Example B control group could be defended as having been selected in exactly the same manner as the treatment group. When "the same manner" is defined as "that manner which will not produce differential regression," however, then the Example B procedure is clearly the correct one.

There would be no problem with either Example A or Example B if the pupils in the treatment and control groups were retested after selection and the second set of scores used as pretest measures. If, on the other hand, the scores used to select the treatment and control group members are used as pretest measures, Example B would be the only suitable selection procedure. If the Example A procedure were used, differential regression would make the Title I treatment appear less effective than it really was.

Deriving the No-Treatment Expectation

Ideally, the treatment and control groups should not differ significantly in terms of initial achievement levels. If the mean pretest scores of the two groups differ by 1 NCE or less, or by 3 percent or less of the observed raw score range, then the mean posttest score of the control group can be used as the no-treatment expectation. However, if the mean pretest scores of the treatment and control groups differ by more than these amounts, then some type of adjustment must be made to the control group's mean posttest score to compensate for the initial pretest difference (assuming that the control group design still appears to be the most valid one to use). Two methods of adjustment are offered--covariance analysis and the

principal-axis or standardized-gain-score method. The choice between these two methods is determined by the nature of the groups.

Covariance analysis should be used if individuals were randomly assigned to the treatment and control groups. In these instances, pretest differences can be attributed to chance sampling fluctuations, and covariance analysis is the appropriate adjustment procedure.

The covariance method of adjustment is also appropriate if existing intact groups are used and assignment to a group, although not random, is believed to be random in effect (Lord, 1967). For assignment to be random in effect, it must be reasonable to assume that the treatment and control groups are, in essence, random samples from a single population. To be considered random in effect, both groups of children must be similar in terms of socioeconomic status, age, sex, and racial composition, and they must attend schools which are similar in size and setting. If group assignment is neither random nor random in effect, applying covariance analysis will systematically underadjust for the pretest difference. To adjust for non-random group differences, the principal-axis method should be used.

The rationales underlying the two adjustment procedures are presented in a separate Technical Paper (No. 12), entitled Statistical Adjustments for Nonequivalent Control Groups. Briefly, covariance analysis assumes that the treatment and control groups were formed through random assignment. The difference between their pretest scores would thus be expected to shrink upon retesting due to statistical regression. The posttest scores are thus adjusted by an amount equal to the pretest difference multiplied by the weighted average of the treatment and control group regression line slopes. The principal-axis or standard-gain-score procedure assumes that the two groups were drawn from different populations and that the difference between pretest

scores would not shrink upon retesting. It adjusts posttest scores by an amount equal to the pretest difference multiplied by the weighted average of the slopes of the two groups' principal axes. This adjustment is larger than the covariance adjustment. Computational procedures are included in the Model B forms and instructions.

NCE Gains

In Model B1 when normed tests are used, the determination of the NCE gain is relatively straight-forward. The posttest percentile of the treatment group and the unadjusted or adjusted posttest percentile of the control group are converted to NCEs. The difference between the two NCE scores is the NCE gain and the measure of project impact.

In Model B2, the determination of the NCE gain is slightly more complex because of the use of a non-normed test. Like Model B1, a gain is computed based on the difference between the treatment group's posttest score and the control group's unadjusted or adjusted posttest score. This difference is measured in terms of raw scores, and cannot be converted directly into NCEs because of the absence of national norms (percentiles).

It can be shown algebraically that NCE gains can be derived by dividing the difference between the observed posttest raw-score mean and the no-treatment expectation by the standard deviation of the national sample and subsequently multiplying it by 21.06. Raw score gains can thus be converted to NCE gains through application of this simple formula.

Unfortunately, non-normed tests have, almost by definition, not been administered to nationally representative samples of children. Thus it is necessary to estimate what the standard deviation would have been had a nationally representative sample been tested. The need to make this estimate is the reason for administering a nationally normed test in conjunction with the normed test in Model B2.

The estimate is based on the assumption that the ratio of the treatment group's standard deviation to the national sample's standard deviation on the normed test is equal to the ratio of the treatment group's standard deviation to the national sample's standard deviation on the non-normed test. Since the two treatment group standard deviations can be calculated from the collected data and the standard deviation of the national sample on the normed test can be obtained from that test's technical manual, the estimated national sample's standard deviation on the non-normed test can easily be derived. The raw score gain divided by this estimate is then multiplied by 21.06, the standard deviation of NCEs. The result is an NCE gain.

Rules for implementation

- Select a control group which is as similar as possible to the treatment group. Be sure that the method of selection will not produce differential regression for the treatment and control groups (or plan on pretesting the group after the selection process has been completed).
- Test the treatment and control groups at the same time on the same day for the pretest and for the posttest. There are no restrictions on testing dates, unless Model A is also implemented.
- Either normed or non-normed tests may be used for evaluation. If a non-normed test is given pre and post, then sometime during the year a normed test must also be given to the treatment group. Only the subtest(s) that cover the same content area as the non-normed test needs to be given.

Implementation Procedures

Step 1: Identify the children who will form the treatment and control groups.

Step 2: Administer the pretest. Both treatment and control groups should be tested on the same day. Make-up tests should be given within two weeks of the original testing date.

Step 3: Score the test and record the scores.

Step 4: For Model B2 only, administer a normed test to the treatment group sometime during the school year so that achievement gains can be converted to NCEs. Only the subtest(s) that covers the subject area to be evaluated needs to be administered.

Step 5: Score the test and record the scores.

Step 6: After the project, administer the posttest. Both treatment and control groups should be tested on the same day. Two-thirds of the project's instruction should occur between the pretest and posttest. Administer make-up tests within two weeks of the original testing date.

Step 7: Score the test and record the scores.

Step 8: Compare the mean pretest scores of the two groups and if necessary adjust the control group's mean posttest score for the initial group difference by using covariance analysis or the principal-axis method.

Step 9: Determine the effectiveness of the project by carrying out the calculations in the Model B1 or B2 instructions.

VI. IMPLEMENTATION OF MODELS C1 AND C2

General Characteristics

This model is less rigorous than Model B, but more rigorous than Model A. Testing costs will be similar to those for Model B and greater than Model A. The calculations needed for Model C are the most complex of all the models. It will be necessary to calculate means, standard deviations, and pretest-posttest correlation coefficients.

Model C1 is a special regression model in which a treatment group and a superior comparison group are formed from an original, intact group based on pretest scores from either a normed or a non-normed test. Both groups are posttested with a normed test. Model C2 follows the same design, but permits pre- and posttesting to be done with a non-normed test. In addition, however, at some point during the school year a normed test must be administered to the treatment group. The standard deviation of the treatment group's normed test scores is calculated and used to convert the gain into units common to all the models, Normal Curve Equivalents.

In Model C, the pretest measure must be used to select the project participants. Those who fall below a selected cutoff score are assigned to the treatment group and the remaining children form the comparison group.

Deriving the No-Treatment Expectation

The no-treatment expectation is derived from the comparison group's post-on-pretest regression line. The rationale for this derivation is as follows. In Figure 3, the positions of 200 children are plotted in terms of their pre- and posttest scores. The children's pretest scores are represented on the horizontal axis and their posttest scores on the vertical axis. The plot forms an ellipse, with more pupils represented in the center of the ellipse than at its edges.

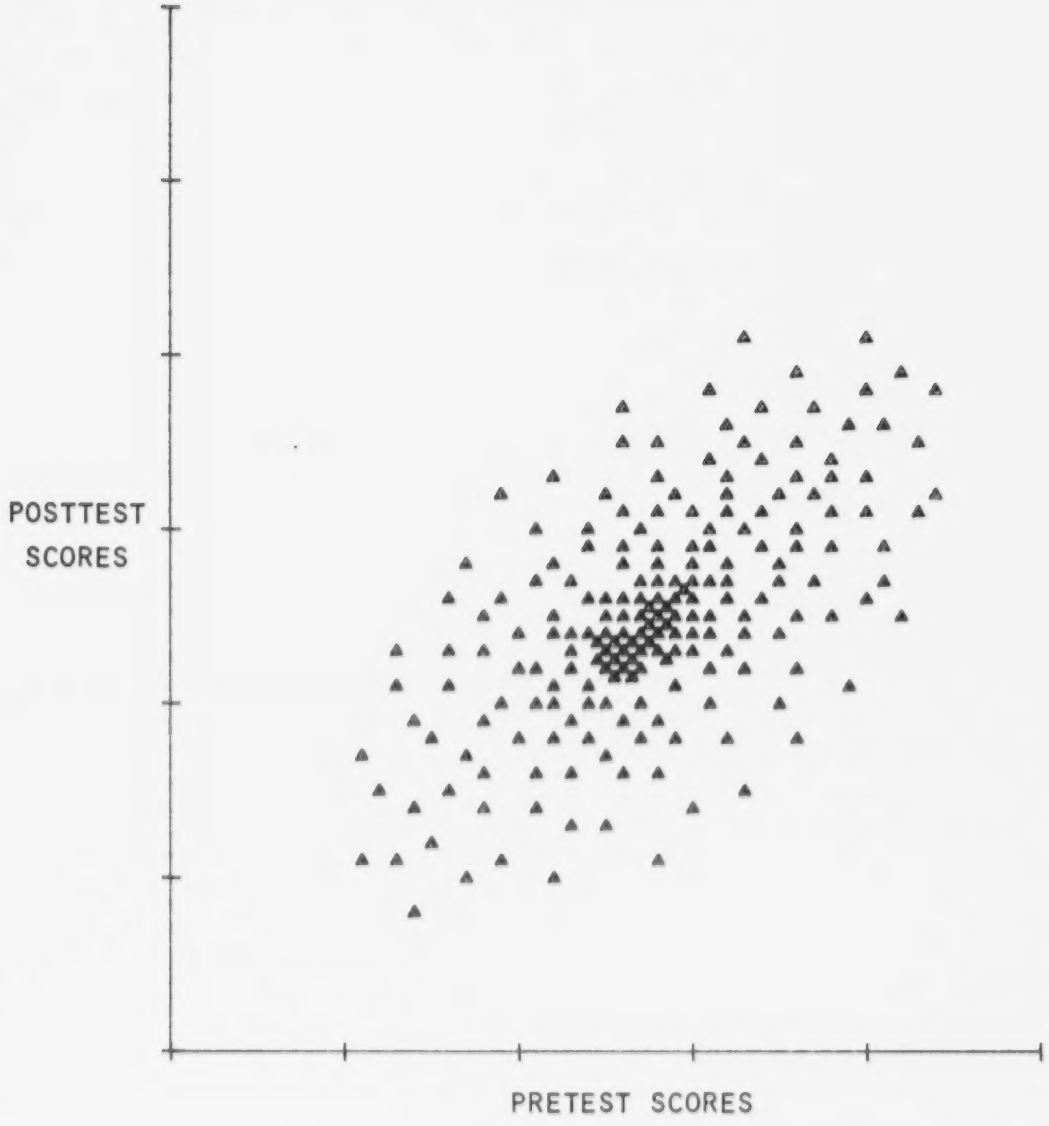

Figure 3. Scatter plot of 200 pairs of pretest and posttest scores.

Notice that students who have the same pretest scores have different posttest scores. For each pretest score, it is possible to find the mean of the corresponding posttest scores. In the ideal situation illustrated in Figure 4, all of these means will fall on a single straight line. This line is the regression line, and it allows us to predict, for any pretest score, the corresponding mean posttest score.

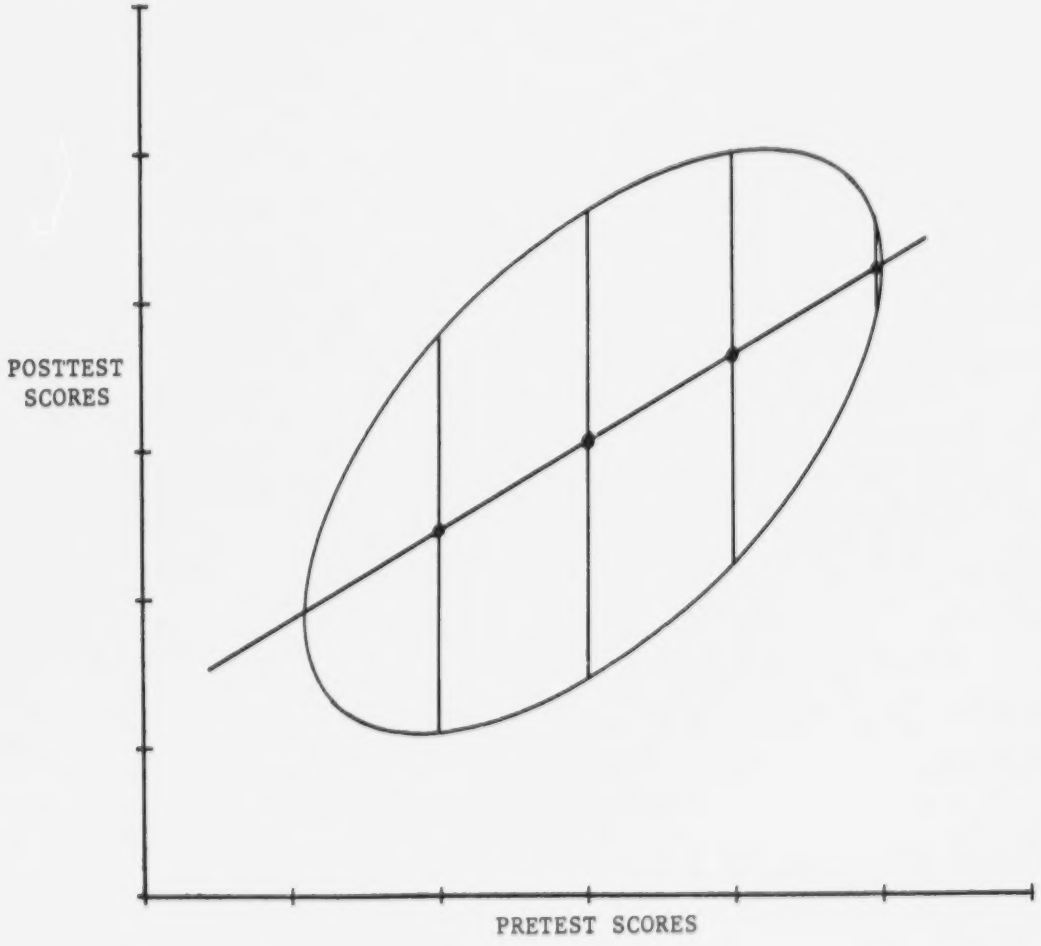

Figure 4. A regression line passing through the midpoints of the posttest score distributions.

In Figure 5, the left-hand portion of the ellipse has been removed. However, since the midpoints of the posttest score distributions corresponding to each pretest score remain unchanged, the truncated ellipse has exactly the same regression line as the full ellipse. Furthermore, the line can be extended in either direction so that mean posttest scores can be predicted for pupils with pretest scores either higher or lower than those represented in the truncated ellipse. It is this feature which enables us to derive a no-treatment expectation from a nonequivalent comparison group. In Model C, a large group of children is pretested. Those scoring below a selected cutoff point participate in the Title I project while those scoring above the point form the comparison group and receive no special treatment. At posttest time, both the treatment and comparison groups are tested again. The scores of both groups can be plotted and, if the treatment was effective, the portion of the plot representing treatment group children should be displaced upward (with respect to the comparison group plot) indicating that the posttest scores were higher than they would have been had the pupils not received the Title I treatment (see Figure 6).

Separate regression lines can be drawn for each group and extended downward or upward. If the treatment group's regression line lies above the downward extension of the comparison group's line, the treatment, presumably, had a positive effect. If the lines are parallel, it can be inferred that the treatment was equally effective for all initial achievement levels in the treatment group. The distance between the two regression lines provides an index of the size of the treatment effect.

NCE Gains

The distance between the regression lines is measured at two different places: where they intersect the cutoff score and where they

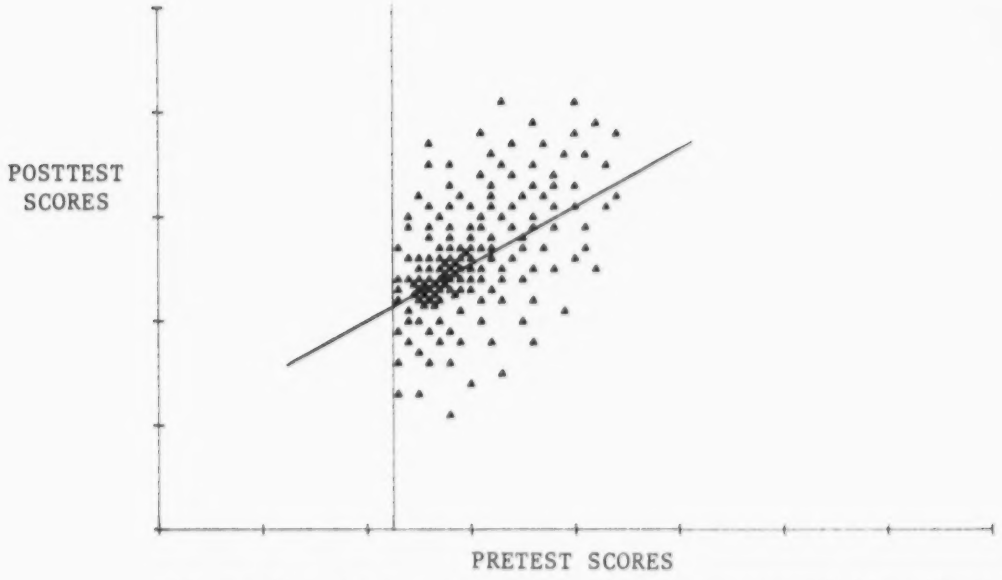

Figure 5. The regression line for a truncated distribution.

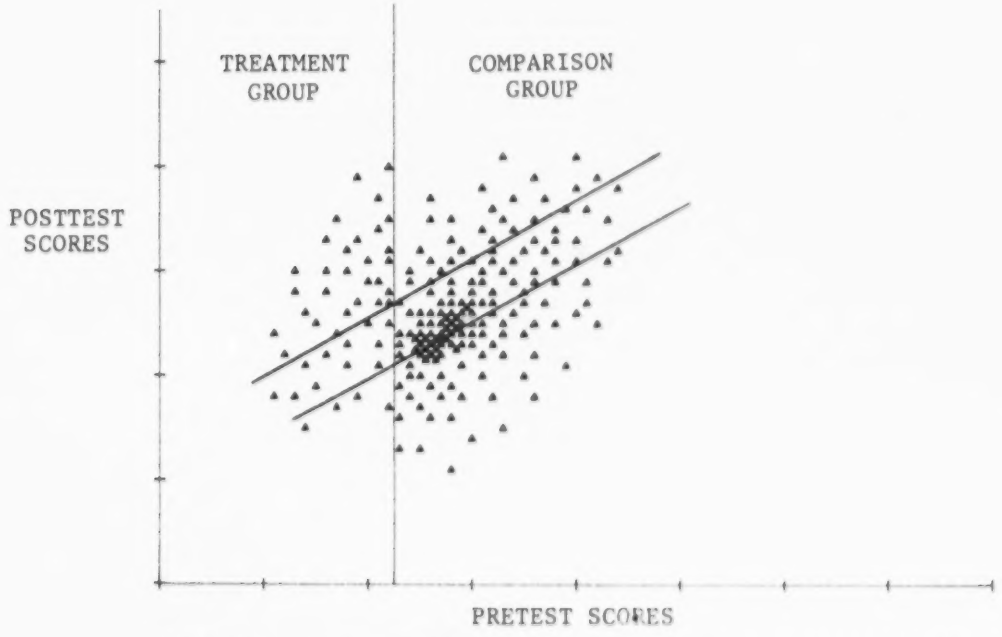

Figure 6. Treatment and comparison group regression lines.

intersect the treatment group's mean pretest score. If the project was equally effective for all children in the treatment, and if neither the test ceiling or floor was encountered, these two distances would be the same.

In Model C1, it is possible to obtain observed and predicted percentile scores at the two places where the gain is measured. The percentiles are converted to NCEs and the NCE gain is computed by subtracting the comparison group value from the treatment group value at each place.

In Model C2, the differences between the regression lines can only be measured in raw score units. Additional computations are required to convert the raw score gains into NCE gains.

It can be shown algebraically that NCE gains can be derived by dividing the difference between the observed posttest raw-score mean and the no-treatment expectation by the standard deviation of the national sample and subsequently multiplying it by 21.06. Raw score gains can thus be converted to NCE gains through application of this simple formula.

Unfortunately, non-normed tests have, almost by definition, not been administered to nationally representative samples of children. Thus it is necessary to estimate what the standard deviation would have been had a nationally representative sample been tested. The need to make this estimate is the reason for administering a nationally normed test in conjunction with the normed test in Model C2.

The estimate is based on the assumption that the ratio of the treatment group's standard deviation to the national sample's standard deviation on the normed test is equal to the ratio of the treatment group's standard deviation to the national sample's standard deviation on the non-normed test. Since the two treatment group standard deviations can be calculated from the collected data and the standard deviation of the national sample on the normed test

can be obtained from that test's technical manual, the estimated national sample's standard deviation on the non-normed test can easily be derived. The raw score gain divided by this estimate is then multiplied by 21.06, the standard deviation of NCEs. The result is an NCE gain.

Rules for Implementation

- A single cutoff score must be used to select treatment group students. Once the cutoff score is established it must be strictly enforced and those students with higher scores cannot later be placed in the treatment.

  The cutoff score does not have to be a single test score. It can be a composite measure composed, for example, of a pretest score and an independently made, numerical, teacher rating of need. As with test results, when using a composite measure a precise cutoff score must be used. A more complete discussion of this topic can be found in Technical Paper No. 10, entitled Composite Scores.

  The pretest measure should correlate highly with the posttest measure, and the higher the correlation the better. A pretest/posttest correlation of at least .4 is required for the comparison group.

- The size of the treatment and comparison groups should be as large as possible. This requirement is especially important in the case of the comparison group, since the stability of its projected regression line is primarily dependent on the number of observations upon which it is based. Model C should probably never be implemented with fewer than thirty children in each group.

- Either normed or non-normed tests can be used. It is not necessary, furthermore, that the same test be given both pre and

post, although a high pretest/posttest correlation is essential. If a non-normed test is given at pre- and posttest times, then a normed test must be administered to the treatment group sometime during the school year. Scores from this test are used to convert the raw score gains derived from the non-normed test into NCEs. Only the subtest that covers the same content area as the non-normed test needs to be given.

- Testing times are not critical in this model. However, all children should be tested on the same day for the pretest and again for the posttest.

- The model assumes that, under no-treatment conditions, the regression of posttest on pretest scores for the total group is homogeneous and linear throughout the entire score range. If it is not, a large apparent treatment effect may be observed where the two regression lines intersect the treatment group's mean pretest score but not where they intersect the cutoff score. This pattern could be observed, even where no treatment was administered.

  If both measures of the treatment effect show positive gains of comparable magnitude, there can be little doubt that the treatment did, in fact, have impact. If, however, a gain is found at the pretest mean but not at the point where the regression lines intersect the cutoff score, the cause could either be a real treatment effect or curvilinear regression, and the latter possibility must be rejected before the former can be accepted.

  Test floor and ceiling effects are the most probable causes of curvilinear regression. To check for floor and ceiling effects, the score distributions for each test should be plotted and examined for excessive skewness. If the scores pile up at one

end or the other, the test may have been too easy or too difficult. An evaluation using these test scores may dramatically overestimate or underestimate the real effectiveness of the project and is likely to produce two very different measures of gain.

A difference between the two measures may also occur because the treatment was more effective for some subgroups served than for others. A project may be designed, for example, so that the amount of instruction each student receives is based on his individual need. Under these circumstances, the project should be most effective with the children farthest below the cutoff, and the gain at the pretest mean should be larger than the one at the cutoff. However, in the absence of this type of specific project-related explanation, the possibility that curvilinear regression might have created the discrepancy between gain measures must be investigated.

The measure of gain at the cutoff score is insensitive to treatment effects which are related to the initial degree of student need. For this reason, it is used only to alert the evaluator to the possibility of curvilinear regression. The measure of gain at the treatment group's mean pretest score is the one used in subsequent aggregation.

Implementation Procedures

Step 1: Administer and score the pretest. The test should be given to all members of the group which will subsequently be divided into treatment and comparison subgroups. The pretest measure should correlate highly with the posttest.

Step 2: If desired, generate a composite score which incorporates the pretest and other measures such as independently made teacher ratings. If composite scores are used, they then become the pretest

measure and all further calculations involving the pretest measure must use the composite scores. In this instance, the composite measure should correlate highly with the posttest.

Step 3: Establish a single cutoff score. Assign all pupils scoring below this value to the treatment group. One convenient way to establish the cutoff score is to determine first how many pupils can be served by the project funds and then to count up from the lowest scoring child until the quota is filled. Once the cutoff is determined, students may not be removed from one group and placed in the other. If exceptions must be made to the group assignment, the scores of these students should not be included in the analysis.

Step 4: For Model C2 only, administer a normed test to the treatment group sometime during the school year so that the achievement gains can be converted to NCEs. Only the subtest(s) covering the subject area to be evaluated need be administered. Score the test and record the scores.

Step 5: Administer and score the posttest. All students in both the Title I group and the comparison group must be posttested. Testing for both groups should occur on the same day. If each of these groups is very large, it is permissible to test a random sample of students from each, but in general this practice is not encouraged.

Step 6: Determine the effectiveness of the project by carrying out the calculations as specified in the Model C1 or C2 instructions.

VII. ADDITIONAL CONSIDERATIONS

Coordinating and Scheduling Activities

Many decisions concerning the evaluation of a project need to be made far in advance of the project's start-up date. The selection of the evaluation design, the development of a testing schedule, and the procedures to be used in identifying the project participants are all important considerations. Teachers, project directors, administrators, and others concerned with the project need to be aware of the evaluation activities since the decisions they make often affect the evaluation process. Close coordination is very important between those responsible for the evaluation and those conducting the project.

Selecting the Model. After reading Chapter 2 and the chapters on model implementation, it should be evident that the selection of an evaluation model must be based on both technical and practical considerations. The choice of the evaluation model should be given a great deal of thought since the model's implementation requirements have major implications for other, related activities.

Establishing a Testing Schedule. In many districts, a large number of different tests are administered to students. Overtesting can be a serious problem. Whenever possible, avoid administering additional tests and take advantage of already existing district-wide or statewide testing programs--assuming, of course, that they meet the model's testing requirements.

Making a chart of the entire testing schedule can be very helpful in avoiding problems. A sample format for such a schedule is shown on the next page. The information should be presented for each project by grade level and for control or comparison groups (if they are used) as well as for the treatment group. The schedule should show the name of the test and the testing dates for the pretest and posttest. (If Model A is selected, be sure that the dates are close

Project Name or Code _____					
Grade	Group (T/C)	Pretest		Posttest	
		Date	Name, Level, Form	Date	Name, Level, Form

Figure 7. Sample format for a testing schedule

to the time at which normative data were collected by the test publisher.) To avoid test floor and ceiling effects, careful thought should be given to the level of tests that are appropriate. Both the level and form of the test selected should appear on the testing schedule. Whenever possible, the level and form should be the same for the pretest and the posttest.

In addition to the testing, the occurrence of other events related to the evaluation needs to be anticipated. The ordering of the tests, for example, must be done well before the first testing date, and the required numbers of each level and form need to be specified. Before testing begins, time should be allocated for training those who are to administer the tests to ensure that everyone follows the same testing procedures. Arrangements must also be made for sending the tests to the proper sites, for collecting answer sheets and booklets after testing has been completed, and for scoring the answer sheets.

Selecting the Students. The models differ in the way in which the project students may be selected. Where the norm-referenced design is used, pretest scores must not be used for selecting participants. The special regression model, on the other hand, requires a strict assignment of pupils to treatment and comparison groups based solely on the pretest measure. Before identifying the Title I participants, review the chosen model's implementation requirements carefully, particularly those relating to the appropriate procedure for selecting students.

Administering the Test. All student testing should be conducted under the same conditions. If a normed test is administered, the directions given to the project students should be the same as those used with the norm group. With Models B and C, primary importance must be placed on making the conditions identical for treatment and control or comparison groups. Refer to Technical Paper No. 13, entitled Collecting Achievement Test Data, for a more complete discussion of this topic.

Analyzing the Data

The essential steps for implementing the models have been described in earlier chapters. Before beginning the basic data analysis, some preliminary screening of the data is often helpful in interpreting the findings.

A. For students with both pre- and posttest scores:

(1) Plot the distribution of the pretest raw scores, standard scores, or NCEs, and compute the mean and standard deviation.

(2) Plot the distribution of the posttest raw scores, standard scores, or NCEs, and compute the mean and standard deviation.

(3) Plot the joint pretest-posttest distribution, and compute the product-moment correlation.

B. For students with pretest scores only:

Plot the distribution of the pretest raw scores, standard scores, or NCEs, and compute the standard deviation.

C. For students with posttest scores only:

These scores are usually not interpretable by themselves, but may be saved for student files or used as baseline data for following-year evaluations.

Calculating the means, standard deviations, and correlations suggested above for students with both pre- and posttest scores is required when deriving the project gain for some of the models. With a little additional effort, score distributions can be plotted and the data from students with only one test score can be reviewed. All of this information is useful when inspecting the data for irregularities and when interpreting the measure of project impact.

Differences between Students with Pre- and Posttest Scores and Those with Pretest Scores Only. Students with incomplete data can differ systematically from students who have all the necessary test scores. Some effort should be made to identify any existing differences. The

lowest scoring students, or students who speak English as a second language, may have dropped out of the project. The highest scoring students may have been graduated from the project. The posttest may have been given at a time of year (e.g., the planting season) when certain groups of children may be absent from school. The existence of any of these factors means that the evaluation data (based on students with all test scores) may not be representative of all the project's participants. As a result, the measure of project gain may be artificially inflated or deflated.

Floor or Ceiling Effects. Skewed distributions, showing many students scoring near the top or bottom of the range, may indicate the presence of floor or ceiling effects. The top of the score range is the highest possible raw score, but a ceiling effect may be encountered even where no student gets a perfect score, as misunderstanding and carelessness will almost always produce a few errors. The bottom of the score range may be zero or, if pupils guessed on a multiple-choice test, it may be the score that would be achieved if students guessed on all the items.

If the treatment children encounter the floor on pretesting, or the ceiling on posttesting, their gains will be underestimated. The impact of the floor or ceiling effect, when it exists, cannot be precisely estimated. However, in a symmetrical distribution, the mean and median will be equal. A skewed distribution will have an impact on the mean but not the median. The median may thus provide an estimate of where the mean would have been if ceiling and floor effects had not occurred.

In instances where the difference between the mean and median is approximately a third of a standard deviation or more, it is likely that the floor or ceiling has been encountered. Where the level of test appears to have been inappropriate, there is no adequate adjustment which can be made to the gain. However, if future evaluations

are conducted with the same test and similar students, administering a different level of the test would certainly be warranted.

Changes in the Standard Deviation. When a change in the size of the standard deviation of test scores occurs between pretest and posttest, the project might be either (a) benefiting the lowest scoring children the most, thus decreasing the range of scores at posttest time, or (b) benefiting the highest scoring children the most, thus increasing the range of scores at posttest time.

A sharp increase in the posttest standard deviation might also indicate that a floor effect existed at pretest time, thereby making the pretest standard deviation smaller than it would have been. Conversely, a sharp decrease in the standard deviation at posttest time might indicate that the group encountered the ceiling of the posttest. The hypothesis that changes in the standard deviations are caused by floor or ceiling effects becomes more credible if the score distributions are skewed (see above section).

Low Correlations. Low correlations between pretest and posttest scores may occur as a result of one or more factors. Usually a low correlation indicates that a test is not measuring the attribute of interest with sufficient reliability. If the skill is not being measured reliably, then project gains will not be measured adequately. Related factors that might cause low correlations are listed below:

- Correlations between test scores may be low if the level of test is changed at posttest time. The two levels may not overlap sufficiently in content and, thus, may be measuring slightly different skills.

- The magnitude of the pretest-posttest correlation is closely related to the grade or age of the children tested. The test scores of younger children will have much lower correlations than those of older children.

- Correlations based on scores from children representing a restricted range of ability will be lower than if a group with a broader range of ability were tested.

- The length of the tests also has an influence on the correlation. The correlation between pre- and posttest scores on a shorter test, e.g., Vocabulary subtest, will be lower than the correlation between scores on a longer test, e.g., Total Reading.

Assumptions Made when NCE Gains Are Averaged

After project gains have been calculated, it is often necessary to present data summaries for larger units such as schools, districts, regions, or States. This requires averaging project or grade-level gains that are based on the use of different test instruments and different evaluation designs. Several basic assumptions are made when NCE gains are aggregated.

Aggregation across models is straightforward since all three measure gains by subtracting an estimate of how well the students would have performed had they not participated in the treatment from a measure of how well they performed after participation. The different methods used in each model to generate the no-treatment expectation enable evaluations to be conducted in a wider variety of conditions than would be possible with any single model. The methods do not produce different kinds of estimates of no-treatment performance, however, and, if it were possible to evaluate a single project by implementing all three models, the three measures of project gain would not differ systematically. This equivalence allows data to be aggregated across models despite there being differences in the procedures used in each.

The assumptions which underlie the aggregation of data from different tests relate to the content coverage of the tests. Clearly, if test data are to be aggregated, the tests must measure at least approximately the same variable or set of variables. There is ample evidence

from the Anchor Test Study (Loret, Seder, Bianchini, & Vale, 1974) that the most widely used standardized reading achievement tests meet this requirement--at least for grades four, five, and six. It is probably safe to assume that other standardized reading tests also "qualify" since visual inspection of their content reveals a high degree of similarity.

There are also data to suggest that at least one criterion-referenced reading test, the Prescriptive Reading Inventory, correlates sufficiently highly with standardized reading achievement tests so that aggregation presents no problem. This test, in fact, was found to correlate as highly with the California Achievement Test (CAT), one of the instruments included in the Anchor Test Study, as alternate forms of the CAT correlated with one another (Roudabush, 1975).

It certainly cannot be expected that all reading tests will intercorrelate as highly as those mentioned above. The correlation between a standardized reading achievement test and a criterion-referenced instrument covering a restricted number of objectives, or that between two criterion-referenced tests encompassing different objectives, will most probably be significantly lower. Even under these circumstances, however, it does not appear unreasonable to aggregate achievement data.

Basically, the assessment of instructional impact is a question of measuring gains rather than status. It follows, then, that gains in any subset of the skills which comprise reading ability will represent an increase in overall ability to read. To use an analogy, it is likely that changes in the circumference of persons' waists may provide nearly as sensitive an index of weight gain or loss as actual weighings. To the extent that this sort of phenomenon holds true with respect to reading ability, it is not necessary to worry excessively about whether one is measuring the universe of skills involved, or a small sample which varies in nearly direct proportion.

Recommended Reading

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080-01460) Also published as Measuring achievement gains in educational projects. Los Altos, CA: RMC Research Corporation, October 1974. (RMC Report No. UR-243)

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Washington, D.C.: U.S. Government Printing Office. (Stock No. 017-080 01516) Also published under the same title in Los Altos, CA: RMC Research Corporation, December, 1975. (RMC Report No. UR-240)

Fourteen brief technical papers have been written which address different topics related to the evaluation and reporting system. Throughout the User's Guide specific papers have been referenced where relevant. The following is a complete list of all the available Technical Papers.

Technical Papers

No. 1 What's Bad About Grade-Equivalent Scores
No. 2 Interpreting NCEs
No. 3 The Regression Effect
No. 4 Selecting a Norm-Referenced Test
No. 5 Characteristics of Eight Commonly Used, Nationally Normed Tests
No. 6 Out-of-Level Testing
No. 7 Local Norms
No. 8 Types of Test Scores
No. 9 Score Conversions
No. 10 Composite Scores
No. 11 Criterion-Referenced Tests
No. 12 Statistical Adjustments for Nonequivalent Control Groups
No. 13 Collecting Achievement Test Data
No. 14 Common Evaluation Hazards

REFERENCES

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1967.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. Anchor test study: Equivalence and norms tables for selected reading achievement tests (grades 4, 5, 6). Office of Education Report 74-305, Washington, D.C.: U. S. Government Printing Office, 1974.

Roudabush, G. E. Estimating normative scores from a criterion-referenced test. Paper presented at the meetings of the American Educational Research Association, Washington, D.C., April 1975.

PREPARED BY RMC RESEARCH CORPORATION
UNDER U.S. OFFICE OF EDUCATION
CONTRACT NO. 300-76-0316